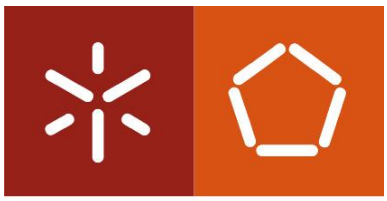


Universidade do Minho
Escola de Engenharia

Alexandre Manuel Da Silva Ribeiro

Implementação de um Sistema de *Business Intelligence* para a análise da Doença Pulmonar Obstrutiva Crónica

Tese de Mestrado
Engenharia e Gestão de Sistemas de Informação

Trabalho efectuado sob a orientação de:
Professora Doutora Maribel Yasmina Santos
Professor Doutor Jorge Manuel Costa da Cruz

Outubro de 2011

Agradecimentos

Um projecto de dissertação é, na maioria do seu tempo, um processo solitário que não poderia ser cumprido sem o apoio de certas entidades e pessoas. Ao concluir este projecto, não posso deixar de lembrar todos os que me apoiaram na sua concretização.

Quero primeiramente agradecer aos meus pais pelo seu permanente apoio incondicional e pela educação que me proporcionaram que fizeram de mim aquilo que eu sou hoje. Todo o apoio dedicado, em todas as vertentes imagináveis, está acima de qualquer gratidão expressa. Apenas o meu amor por eles alguma vez os poderá compensar.

À minha orientadora, Professora Doutora Maribel Yasmina Santos, pela belíssima orientação, dedicação, apoio e amizade, estando constantemente disponível para ouvir as minhas dúvidas e tecer palavras de incentivo, agradecendo também os bons comentários e sugestões que me facultou durante todas as fases da dissertação.

Ao meu co-orientador, Professor Doutor Jorge Manuel Costa da Cruz, da Faculdade de Medicina de Lisboa, pela sua orientação a nível clínico, fundamental para a compreensão da problemática e para a validação dos resultados obtidos.

Ao Dr. Artur Teles de Araújo, presidente da Fundação Portuguesa do Pulmão, pela disponibilidade e pelos dados concedidos indispensáveis à realização deste projecto de dissertação.

Aos meus professores do curso de Mestrado em Engenharia e Gestão de Sistemas de Informação, pelos conhecimentos transmitidos e dedicação na sua tarefa de formar mestres.

À minha irmã, Elsa, pela amizade, incentivo, apoio e paciência durante estes anos de mestrado.

À Marlene, minha melhor amiga e eterna namorada, pelo orgulho que sente por mim, por não me ter cobrado as minhas ausências, por me ter dado a mão em todas as situações que precisei e não precisei e por me fazer acreditar de que era capaz.

A todos os meus amigos, pelo apoio e incentivo incondicional, desde aqueles mais próximos até aqueles mais distantes, os meus sinceros agradecimentos.

Resumo

O aparecimento dos sistemas de *Business Intelligence* e das tecnologias associadas foi influenciado pela necessidade das organizações em processar a informação de uma forma cada vez mais eficiente de maneira a suportar a tomada de decisão nas organizações. De entre o vasto conjunto de áreas aplicacionais destaca-se neste trabalho a área da Saúde, que apresenta necessidades ao nível do armazenamento e análise de dados de forma a disponibilizar informação relevante aos seus processos de tomada de decisão.

Apenas com informação acerca dos padrões de incidência e das possíveis causas será possível combater eficazmente uma doença e aplicar políticas de saúde mais adequadas. O estudo da Doença Pulmonar Obstrutiva Crónica, apresentado neste trabalho de dissertação e resultante do trabalho de investigação e desenvolvimento que tem vindo a ser realizado entre o Departamento de Sistemas de Informação da Universidade do Minho e a Fundação Portuguesa do Pulmão que não tem precedentes a nível nacional, pode ser o princípio de um contributo para a formulação de novo conhecimento neste ramo da medicina.

Dada a necessidade de avaliar de uma forma contínua dados sobre esta doença respiratória, este trabalho propõe uma solução que passou pela concepção e implementação de um sistema de *Business Intelligence* que vai auxiliar a tomada de decisão dos profissionais de saúde desta área, graças à disponibilização de mecanismos de recolha, exploração e análise de dados. O sistema engloba uma aplicação *Web* e a sua respectiva base de dados operacional com o intuito de gerir e armazenar os dados recolhidos pela Fundação Portuguesa do Pulmão e um *Data Warehouse* que serve de suporte para as técnicas avançadas de análise de dados como os cubos *On-Line Analytical Processing* ou os algoritmos de *Data Mining*, que permitiram a identificação de padrões de incidência e de modelos preditivos da Doença Pulmonar Obstrutiva Crónica.

Os resultados obtidos permitem verificar a utilidade e a adequabilidade de todos os componentes do sistema de *Business Intelligence* implementado. A integração dos vários componentes num único sistema, incluindo a aplicação *Web* e o modelo do *Data Warehouse*, assim como as técnicas utilizadas para a análise de dados, permitiram confirmar alguns factores de risco relacionados com a Doença Pulmonar Obstrutiva Crónica e identificar padrões desta doença, assim como construir modelos preditivos. No entanto, por outro lado, os resultados obtidos também vêm reforçar a ideia de que a doença é difícil de diagnosticar e que uma boa prevenção, realizando os exames específicos periodicamente, é fundamental.

Abstract

The emergence of Business Intelligence systems and related technologies was influenced by the need for organizations to process information in an increasingly efficient way in order to support decision making in organizations. Among the wide range of applications areas, this work focuses its attention in the health care domain. This application area has storage and data analysis requirements in order to make available valuable information to the decision making process.

It will only be possible to effectively combat a disease and apply the most appropriate health policies with information about the patterns of incidence and possible causes of a disease. The study of the Chronic Obstructive Pulmonary Disease, presented in this work, is the result from the research and development work that has been held between the Department of Information Systems at the University of Minho and the Portuguese Lung Foundation, which has no precedent in Portugal. This work can contribute to the development of new knowledge in this branch of medicine.

Given the need to continuously analyse data of this respiratory disease, this study proposes a solution that includes the design and implementation of a business intelligence system that will assist the decision making of health professionals in this area, making available data collection, exploration and analysis mechanisms. The system includes a Web application and its respective operational database in order to manage and store data collected by the Portuguese Lung Foundation, and a data warehouse that serves as support for the advanced techniques of data analysis such as Online line Analytical Processing cubes and Data Mining algorithms, which allowed the identification of patterns of incidence and predictive models of Chronic Obstructive Pulmonary Disease.

The results obtained allow us to verify the utility and the suitability of all the business intelligence system components implemented. The integration of the several components in a single system, including the web application and the data warehouse model, as well as the techniques used to analyze data, allowed confirming some risk factors related to Chronic Obstructive Pulmonary Disease and identifying patterns of this disease, as well as building predictive models. However, on the other hand, the obtained results also reinforce the idea that the disease is difficult to diagnose and that a good prevention, doing specific tests periodically is essential.

Índice

Índice de Figuras

Índice de Tabelas

Acrónimos e Abreviaturas

AD	Árvores de Decisão
BDO	Base de Dados Operacional
BI	*Business Intelligence*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
DCBD	Descoberta de Conhecimento em Base de Dados
DER	Diagrama de Entiades e Relacionamentos
DM	*Data Mining*
DPOC	Doença Pulmonar Obstrutiva Crónica
DSA	*Data Staging Area*
DW	*Data Warehouse*
EIS	*Executive Information System*
ETL	*Extraction, Transformation, Loading*
FEF25-75	*Forced Expiration Flow 25% - 75%*
FEV1	*Forced Expiration Volume in one second*
FPP	Fundação Portuguesa do Pulmão
FVC	*Forced Vital Capacity*
GOLD	*Global Initiative for Chronic Obstructive Lung Disease*
HOLAP	*Hybrid On-Line Analytical Processing*
IR	Indução de regras
KPI	*Key Performance Indicators*
MOLAP	*Multidimensional On-Line Analytical Processing*
OLAP	*On-Line Analytical Processing*
OLTP	*On-Line Transactional Processing*
ONDR	Observatório Nacional Das Doenças Respiratórias
RNA	Redes Neuronais Artificiais
ROLAP	*Relational On-Line Analytical Processing*
SAD	Sistema de Apoio à Decisão
TI	Tecnologias de Informação

Capítulo 1 – Introdução

Neste capítulo é descrito o enquadramento do projecto de dissertação e é referida a motivação para a realização do mesmo. Também é descrita a finalidade da dissertação e os seus principais objectivos, assim como as metodologias de investigação utilizadas. Conclui-se o capítulo com a apresentação da estrutura do presente documento.

1.1 Enquadramento e Motivação

Este projecto de dissertação está enquadrado no trabalho de investigação e desenvolvimento que tem vindo a ser realizado entre o Departamento de Sistemas de Informação da Universidade do Minho e a Fundação Portuguesa do Pulmão (FPP).

Como se trata de um problema real de uma organização fundada recentemente, este projecto de dissertação pretende ser um contributo inicial valioso para este ramo da medicina em Portugal, mais especificamente para a FPP. Isto porque, como se trata de um estudo novo e que não tem precedentes a nível nacional, poderá trazer informação útil e contribuir para a formulação de novo conhecimento na área.

Apenas com informação acerca dos padrões de incidência e das possíveis causas será possível combater a doença e aplicar melhores políticas de saúde, e como a Doença Pulmonar Obstrutiva Crónica (DPOC) é uma doença muito ligada a comportamentos, é possível intervir. Uma das lacunas actuais da sociedade portuguesa no ramo da medicina encontra-se na promoção da saúde respiratória[1]. Se tivermos em conta, por exemplo, toda a prevenção feita em torno das doenças cardiovasculares ao longo destes últimos anos, mas sobretudo toda a sensibilização realizada, nomeadamente em questões como o colesterol ou o controlo da tensão arterial, verifica-se que a incidência dessas doenças tem diminuído.

[1] Entrevista ao Dr. Teles de Araújo, dia 29 de Outubro de 2009, Sic Notícias, Edição da Manhã.

A recolha de dados dos utentes numa simples folha de cálculo, muitas vezes com erros e com dados incompletos, não permite uma análise precisa e profunda da informação recolhida. Outro problema existente nesta área é a falta de pneumologistas em Portugal (ONDR, 2009), encontrando-se apenas 498 especialistas inscritos na Ordem do Médicos, o que equivale a 1 pneumologista por 21 283 habitantes, número claramente inferior a outras especialidades, como a Cardiologia (1 / 13 819) ou a Pediatria (1 / 7 260). Esta falta de especialistas e a falta de meios informáticos para a recolha e análise de dados fazem com que seja muito difícil, senão impossível, definir a quem tem de ser dirigida a prevenção desta doença, mas mais importante ainda, como tem de ser feita.

Segundo os dados mais recentes do Observatório Nacional Das Doenças Respiratórias (ONDR), a prevalência de DPOC no Mundo é de 63,6 milhões de doentes, dos quais 11,3 milhões se encontram na Europa. A DPOC é responsável por 3 milhões de mortes anuais, o que a coloca no 4º lugar como causa de morte, responsável por 5,36% dos óbitos. Nos países de alto rendimento *per capita* (nos quais se inclui Portugal), a DPOC é responsável por 3,4% dos óbitos, sendo a 5ª causa de morte, atrás da doença isquémica coronária, das doenças cérebro-vasculares, do cancro do pulmão e das infecções das vias aéreas inferiores. A DPOC é a 5º causa de incapacidade, a partir dos 60 anos, nos países de alto rendimento *per capita* e a 7ª causa nos grupos etários entre os 0 e os 59 anos.

Em Portugal, os internamentos por DPOC constituem a 2ª causa de internamento por doença respiratória e apresentam grande variabilidade de ano para ano. Em 2008 foram internados 9301 doentes com o diagnóstico principal de DPOC. Entre 2002 e 2006 a mortalidade geral por DPOC aumentou 5,7%. Apesar de haver variações de ano para ano, os internamentos hospitalares por agudização da DPOC têm vindo a aumentar e praticamente duplicaram desde 1994. Os dados do Instituto Nacional de Estatística colocam a DPOC como a 5ª causa de morte por doença em Portugal (2682 óbitos em 2006), atrás das doenças cardiovasculares, diabetes, pneumonias e cancro do pulmão.

No entanto, a prevalência da DPOC em Portugal é subestimada. Um estudo apresentado no 26º Congresso de Pneumologia que se realizou em Dezembro de 2010 revelou que a prevalência desta doença crónica passou de 5,3 % - número de um estudo anterior efectuado pela Sociedade Portuguesa de Pneumologia - para 14,2%. Esses dados são o resultado do projecto *Global Initiative for Chronic Obstructive Lung Disease* (GOLD) em Portugal e foram apresentados pela primeira vez a nível nacional. A investigação procurou apresentar uma

estimativa da prevalência da DPOC em adultos com 40 ou mais anos, num universo de 2 700 000 habitantes da cidade de Lisboa. Todavia, é possível que, por exemplo, no Norte do país este problema seja ainda mais premente, isto porque é mais frequente nessa região do país haver uma grande exposição à combustão de biomassas, sobretudo madeiras, em locais fechados através de fogões de sala e em cozinhas. Essa omissão de dados poderá significar que essa prevalência de 14,2% da DPOC em Portugal seja, uma vez mais, subestimada.

São, então, vários os indicadores que apontam no sentido da necessidade de uma apropriada caracterização epidemiológica da situação em Portugal. Trata-se de uma doença crónica e o acompanhamento destes doentes obriga a um acompanhamento de proximidade. A DPOC exige que o seu diagnóstico e a sua caracterização devem ser os mais precoces possíveis, pois leva à incapacidade de longa duração e, em casos extremos, à morte.

Dada a relevância do estudo da DPOC pelas razões evidenciadas acima e dada a necessidade de analisar de forma contínua dados sobre a incidência desta doença, juntando o facto da falta de ferramentas que auxiliem os profissionais de saúde nesta tarefa (ONDR, 2011), este projecto de dissertação pretende propor uma solução que passa pela concepção e implementação de um sistema de *Business Intelligence* (BI) que vai auxiliar a tomada de decisão, disponibilizando mecanismos de recolha, exploração e análise de dados. Tal permitirá identificar e suportar a aplicação de políticas de saúde da DPOC em Portugal, ajudando a caracterizar de uma forma mais eficaz e mais eficiente esta doença respiratória. Pretende-se com este projecto propor um sistema que, além da análise de dados, permita ainda a identificação de padrões de incidência e modelos preditivos.

Este projecto engloba ainda a concepção e implementação de uma aplicação *Web* e da sua respectiva base de dados operacional, para a gestão e para o armazenamento dos dados recolhidos pela FPP e tratará de todo o processo de transição destes dados para um *Data Warehouse* (DW), que servirá de suporte à aplicação de técnicas avançadas de análise de dados como os cubos *On-Line Analytical Processing* (OLAP) e os algoritmos de *Data Mining* (DM).

A melhor maneira de combater uma doença é ter o maior conhecimento possível sobre a mesma.

1.2 Finalidade e Objectivos do Trabalho

A finalidade deste projecto de dissertação é o de conceber e implementar um sistema de BI que permita a análise de dados da FPP e que possibilite a identificação de padrões de incidência e de modelos preditivos de Doenças Obstrutivas das Vias Aéreas, mais precisamente da DPOC.

Para concretizar este projecto de dissertação, é necessária a realização de quatro objectivos, nomeadamente:

1. Concepção e implementação de uma aplicação *Web* e da sua respectiva base de dados operacional para a introdução dos dados dos utentes e para a gestão dos mesmos;
2. Definição da arquitectura do sistema de BI a implementar, caracterizando os seus principais componentes e as tecnologias a utilizar;
3. Definição da arquitectura (modelo de dados) do DW, que servirá de suporte ao armazenamento de dados e proceder à sua respectiva implementação;
4. Análise dos dados recorrendo a tecnologias e técnicas avançadas como o OLAP e os algoritmos de DM para identificação de padrões e tendências nos dados.

Com a realização dos objectivos apresentados em epígrafe, espera-se que seja implementado um sistema de BI que permita a identificação de padrões de incidência e modelos preditivos da DPOC que possa trazer informação útil e contribuir para a formulação de um novo conhecimento na área.

Assim, como principais resultados, destacam-se a definição da arquitectura do sistema de BI e a sua posterior implementação. Esta implementação passa, não só pela concretização dos repositórios de dados, mas também pela análise dos dados armazenados.

Paralelamente, e uma vez que a FPP não possui uma aplicação para a recolha dos dados, será disponibilizada uma plataforma *Web* para a recolha dos dados e seu armazenamento, a qual será completamente integrada com o sistema de BI proposto.

Para além destes resultados, foi escrito um artigo científico que foi submetido e aceite na conferência "*The First International Conference on Business Intelligence and Technology*". O artigo inclui a definição de toda a arquitectura do sistema e dos seus repositórios e apresenta os primeiros resultados sobre o trabalho realizado (Ribeiro, Dinis, & Santos, 2011).

1.3 Metodologia de Investigação

A primeira metodologia de investigação utilizada foi, naturalmente, a revisão bibliográfica. Esta serviu essencialmente para enquadrar os conceitos subjacentes a este projecto de dissertação e para verificar o estado da arte neste domínio.

Foi necessária a definição da arquitectura do sistema de BI a implementar, assim como a definição do modelo de dados que guiou a implementação do DW do sistema. Após esses passos, foi essencial descrever os mecanismos de extracção, transformação e carregamento dos dados recolhidos pela FPP e analisar esses mesmos dados recorrendo a sistemas de processamento analítico de dados e a algoritmos de DM.

Para além deste sistema de BI que tem como base dados históricos da FPP, e como outro contributo futuro para a FPP, foi também desenvolvida uma aplicação *Web* (*interface*) que integra este sistema de BI e que permite a inserção, a gestão e o armazenamento dos dados dos futuros inquéritos da FPP numa base de dados operacional em vez da tradicional recolha de dados (sócio-económicos e medições do exame de espirometria) do utente em papel e posterior registo numa simples folha de cálculo, muitas vezes com erros e com dados incompletos que não permitem uma análise precisa e profunda da informação recolhida. Na implementação da base de dados operacional foram consideradas questões de confidencialidade dos dados dos utentes, nomeadamente naqueles que permitem a sua identificação.

Tanto para o sistema de BI propriamente dito, como para a aplicação *Web*, foi usada a metodologia *Constructive Research* (ou *Design Research*), isto porque esta abordagem aplica-se à realização de estruturas inovadoras que são destinadas a resolver problemas do mundo real e trazer uma contribuição à teoria de uma determinada área e portanto aplica-se a este estudo.

Para além das metodologias já referidas, foi usada a prova de conceito. Esta última foi utilizada para validar o sistema de BI que foi proposto. O sistema apresentado tem uma determinada arquitectura e componentes que foram implementados através de um protótipo. A concretização do protótipo evidencia a exequibilidade do sistema proposto neste trabalho.

No que diz respeito à técnica de recolha dos dados, os dados que serão utilizados para o estudo foram disponibilizados pela própria FPP. Quanto à análise dos dados recorrendo a sistemas de processamento analítico de dados, e.g. cubos OLAP, e à identificação de modelos preditivos e padrões associados a esses dados recorrendo a algoritmos de DM, foi usada a metodologia *Quantitative Positivist Research* e o método *Archival Research*, porque foram estudados os dados históricos da FPP e não dados novos.

Como referido acima, os dados que foram estudados já se encontram recolhidos e portanto não foi feito nenhum *survey* novo para obter os dados necessários à realização da investigação. Porém, para tornar o sistema de BI mais completo, foi implementada uma aplicação *Web* que permitirá as futuras recolhas de dados dos utentes através de um questionário, adaptado para o efeito, cujos dados serão enviados para uma base de dados operacional fazendo com que os dados sejam armazenados de uma forma uniforme e facilitando assim os processos seguintes do sistema de BI.

1.4 Estrutura do Documento

A dissertação é constituída por um total de cinco capítulos e dois anexos. O primeiro capítulo tem como propósito contextualizar o projecto, descrevendo alguns dos problemas inerentes assim como a sua finalidade e principais objectivos. As metodologias de investigação usadas são também apresentadas e justificadas neste primeiro capítulo.

O segundo capítulo retrata a revisão de literatura e os principais conceitos, fundamentos e metodologias que alicerçaram todo o trabalho desenvolvido. É dada ênfase aos tópicos mais relevantes para o projecto.

O sistema de BI concebido e desenvolvido é apresentado no capítulo três, começando pela caracterização dos dados disponíveis. Também neste capítulo é descrita a arquitectura do sistema assim como a sua implementação.

No quarto capítulo é descrito o estudo da incidência da DPOC, tendo como base os dados disponíveis, recorrendo em primeiro lugar à componente OLAP e depois à componente de DM. Ao longo do capítulo, evidenciam-se os resultados obtidos para cada uma das abordagens.

Finalmente, as conclusões do projecto ficam remetidas para o quinto capítulo, onde serão apresentadas uma síntese do trabalho realizado, assim como as principais contribuições deste projecto, mas também algumas propostas de trabalho futuro.

Capítulo 2 – Enquadramento Conceptual e Tecnológico

Neste capítulo é feito o enquadramento teórico da dissertação. São retratados a revisão de literatura e os principais conceitos, os fundamentos e as metodologias que alicerçaram todo o trabalho desenvolvido. Após a familiarização dos conceitos e das características que qualificam os sistemas de BI, são dadas a conhecer as principais tecnologias e técnicas de suporte que serão utilizadas no âmbito desta dissertação. São elas os sistemas de *Data Warehousing*, os sistemas de processamento analítico de dados e o DM. Como o estudo incide sobre a DPOC, esta doença não podia deixar de ser retratada neste capítulo. É feita uma breve apresentação da doença, assim como a descrição dos seus principais factores de risco, das formas de ser diagnosticada e dos seus níveis de gravidade.

2.1 Os Sistemas de *Business Intelligence*

Na conjuntura actual, o ambiente organizacional está em constante mudança e está a tornar-se cada vez mais complexo. As organizações, sejam públicas ou privadas, estão sobre constantes pressões o que as obriga a ter a necessidade de responder rapidamente às condições de mudança e a ser inovadoras no modo de operarem. Tais actividades requerem que a organização seja ágil e que tome frequentemente rápidas decisões (muitas das vezes de alguma complexidade), sejam elas estratégicas, tácticas ou operacionais. Portanto, são requeridos uma enorme quantidade de dados, de informação e de conhecimento para tomar essas decisões.

Assim, a organização, ao actuar num mundo global, está num estado permanente de necessidade de informação e conhecimento, pelo que estes constituem o alicerce de uma organização e são componentes fundamentais e indispensáveis à sua existência. Esse facto leva à aceitação por parte dos gestores da organização de que quem dispõe de informação (suporte para o conhecimento) de boa qualidade, fidedigna, em quantidade adequada e no

momento certo, adquire vantagens competitivas. Pelo contrário, a falta de informação dá muitas vezes azo a erros e à perda de oportunidades.

Com o aumento da concorrência e da complexidade do meio ambiente, existe a necessidade de, no mundo empresarial, obter melhores recursos do que os seus concorrentes e de optimizar o seu uso. Para processar estes recursos e para tomar a decisão adequada de uma forma atempada, são então necessários suportes computorizados. Esta necessidade por parte das empresas veio influenciar o aparecimento de conceitos como o BI e das suas tecnologias associadas.

2.1.1 Conceitos e características

O BI é um conceito que apareceu pela mão do *Gartner Group,* em meados da década de 90. Contudo, este conceito de inteligência empresarial já era usado pelos sistemas existentes na década de 70. Nesse tempo, os sistemas de *reports* eram estáticos e não possuíam a capacidade de elaborar análises sobre um determinado conjunto de informação. Na década de 80 começou a haver uma evolução acelerada da tecnologia e uma consequente crescente mudança organizacional. Com essas mudanças emergiu o conceito de *Executive Information System* (EIS). A evolução tecnológica começou a permitir *reporting* multidimensional, previsão e análises com recurso a informação organizacional. Os gestores começaram a poder tomar decisões com fundamento informacional e não intuitivo. Na década de 90, começaram a ser comercializados produtos tecnológicos que possuíam estas características e eram chamadas de ferramentas de BI (Turban, Sharda, & Delen, 2010).

Alguns autores afirmam mesmo que o BI não é mais do que um termo actual para os Sistemas de Apoio à Decisão (SAD) orientados aos dados (Alter, 1999). As evoluções associadas ao BI datam de iniciativas de 1985 de criação de um SAD que fosse capaz de interligar informações de vendas e os dados do *scanner* de um retalho (Pareek, 2006). Ainda segundo a ideologia de Alter, pode-se concluir que o BI substituiu os SAD orientados aos dados porque muitas abordagens desenvolvidas para o auxílio da tomada de decisão incluem o OLAP e o DM. O OLAP surgiu por causa das dificuldades associadas à análise dos dados numa base de dados operacional (BDO) exposta a contínuos *updates*. Então, para não abrandar esses processos continuados, os processos analíticos acedem a outra base de dados desenhada especificamente para suportar essas análises, a qual é carregada periodicamente, o DW. É a

base sobre a qual o OLAP e o DM podem ser aplicados sem perturbar os processos operacionais. Por outras palavras, estes componentes essenciais do BI estão a substituir a forma antiga de executar o SAD orientado aos dados. Os conceitos associados ao DW, ao OLAP e ao DM serão apresentados nas próximas subsecções.

Os sistemas BI são definidos como ferramentas analíticas que têm o objectivo de analisar dados organizacionais para posteriormente disponibilizar informação aos gestores, de forma a melhorar a tomada de decisão (Cody, Kreulen, Krishna, & Spangler, 2002).

Actualmente, os sistemas de BI possuem inteligência artificial de forma a poderem prever tendências futuras, ajudando os gestores numa tomada de decisão muito mais eficaz. (Turban, et al., 2010)

O BI pode também ser definido como um *umbrella term* que combina arquitecturas, ferramentas, base de dados, ferramentas analíticas, aplicações e metodologias (Turban, et al., 2010). Segundo o mesmo autor, o grande objectivo do BI é permitir o acesso e manipulação interactiva dos dados, por vezes em tempo real, de forma a dar uma perspectiva aos gestores e analistas para a tomada de decisão.

As tarefas que normalmente estão associadas aos sistemas de BI são (Santos & Ramos 2006):

- Elaborar previsões com base em informação histórica da organização;
- Criar cenários para verificar o impacto da alteração de determinadas variáveis;
- Permitir acesso *ad-hoc* aos dados para responder a questões que não se encontram pré-definidas;
- Conhecer mais profundamente a organização.

Na Figura 2.1 pode-se observar uma arquitectura de alto nível de um sistema de BI. Esta arquitectura baseia-se numa arquitectura de infra-estrutura tecnológica de apoio ao BI proposta por Han e Kamber (2001) e noutra arquitectura de BI apresentada por Eckerson (2003b). A arquitectura exposta é constituída por três níveis: o ambiente *Data Sources*, o ambiente DW e o ambiente *Business Analytics*.

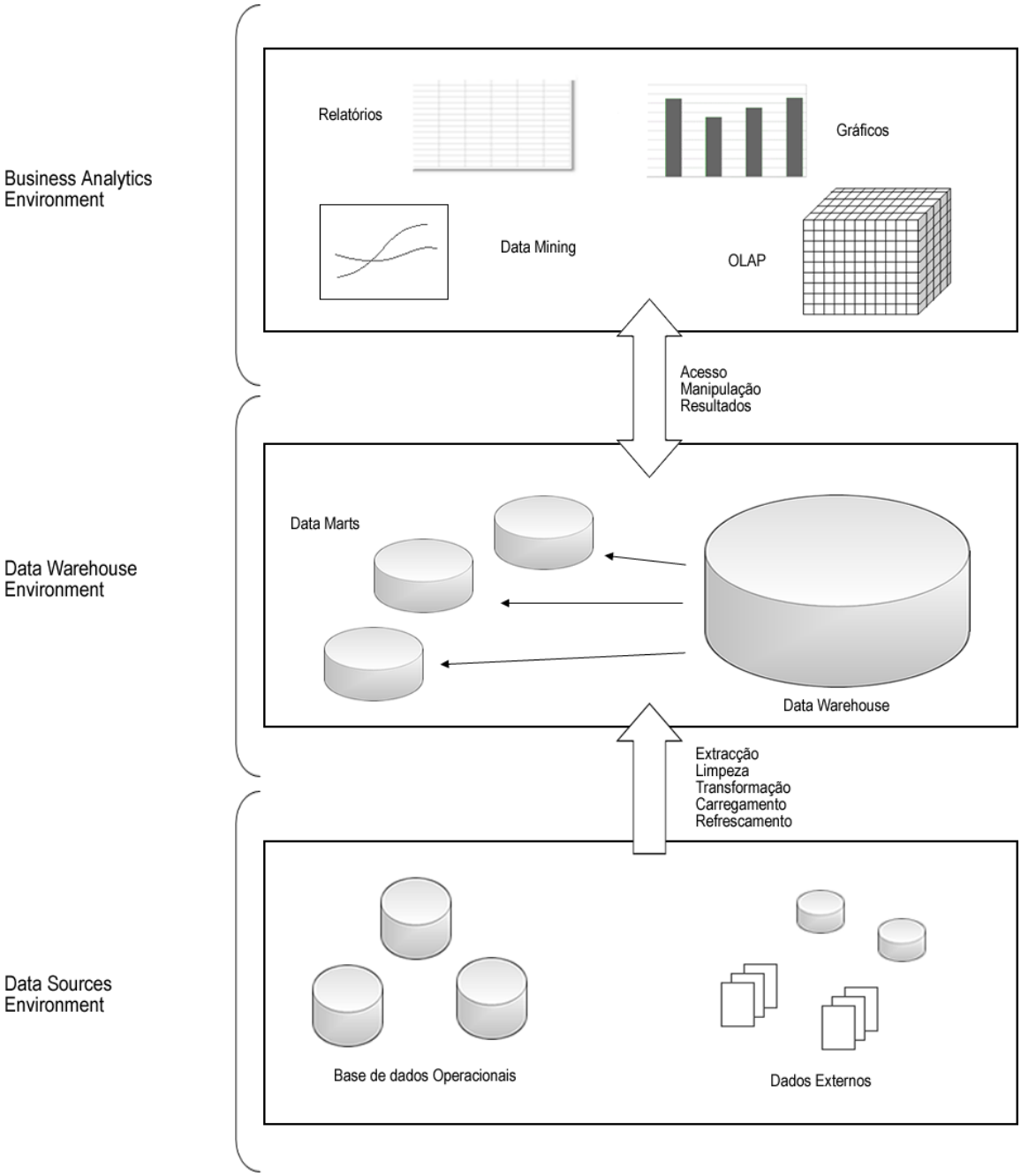


Figura 2.1 - Arquitectura de alto nível de um sistema de *Business Intelligence*

Os três níveis da arquitectura apresentada na Figura 2.1 são de seguida muito brevemente descritos:

- Um dos níveis representa o ambiente das *Data Sources*, no qual estão integradas todas as fontes de dados possíveis para a construção do DW, desde as BDO e de

outras bases de dados da organização, até às bases de dados externas da organização.

- O nível seguinte é o ambiente DW. Integra o DW da organização e/ou os seus diversos *Data Marts.* Estes, DW ou *Data Marts*, são carregados a partir das várias fontes de dados existentes através de ferramentas de *Extraction, Transformation, Loading* (ETL).
- O último nível representa o ambiente *Business Analytics*. Neste é possível trabalhar os dados, acedendo ao DW, utilizando uma variedade de técnicas e ferramentas, como o OLAP e o DM, que serão abordados mais detalhadamente nas próximas subsecções.

2.1.2 As Tecnologias de Suporte

Para além dos conceitos apresentados anteriormente, existem ainda um conjunto vasto de tecnologias que podem estar associadas ao BI, como por exemplo as folhas de cálculo, os próprios SAD, os *Geographical Information Systems*, os *Scorecards*, ou os *Dashboards*. Apesar disso, neste projecto de dissertação apenas serão apresentados os conceitos associados aos sistemas de *Data Warehousing*, aos sistemas de processamento analítico de dados e ao DM, uma vez que são as tecnologias utilizadas no âmbito desta dissertação.

2.1.2.1 Os Sistemas de *Data Warehousing*

Um DW é um repositório de dados construído especificamente para suportar a tomada de decisão da organização. Os dados são estruturados de maneira a estarem disponíveis num formato válido e consistente para permitir actividades de processamento analítico, como consultas já pré-definidas ou consultas *ad-hoc*, e a elaboração de *reports* com os *Key Performance Indicators* (KPI) da organização. Também é possível explorar os dados armazenados no DW recorrendo a técnicas de DM. Estas técnicas integram conceitos provenientes de várias áreas, como por exemplo da inteligência artificial, da aprendizagem automática e da estatística, e que permite identificar tendências ou padrões nos dados que de outra maneira muto dificilmente seriam descobertos. Os conceitos associados ao DM são abordados mais detalhadamente na subsecção 2.1.2.3.

É usual introduzir o conceito de DW referindo as suas características fundamentais. Inmon, que empregou o termo pela primeira vez em 1991, define um DW como um conjunto de dados integrado, orientado por assunto, não volátil e estruturado temporalmente de maneira a suportar os gestores no seu processo de tomada de decisão (Inmon, 2005). Importa então clarificar estes quatro atributos da definição de Inmon:

- Orientado por assunto: Os dados são organizados por assuntos, como por exemplo as vendas, os produtos ou os clientes. Esta orientação aos principais assuntos da organização providencia uma visão simples de um determinado assunto e permite a análise de dados para suportar o processo de tomada de decisão. Ao contrário das BDO, que estão mais orientadas para as operações do dia-a-dia;

- Integrado: A integração está estreitamente relacionada com a orientação por assuntos. Os DW são construídos a partir de diversas fontes com os mais diversos formatos. Portanto, os DW têm de ter a informação proveniente destes últimos armazenada de uma forma consistente. Esta consistência é obtida graças a técnicas de limpeza e integração que resolvem os conflitos e discrepâncias dos nomes dos atributos, das unidades de medidas, entre outros. Presume-se então que um DW tem de estar totalmente integrado;

- Estruturado temporalmente: Um DW mantém dados históricos. Os dados não fornecem necessariamente o estado actual da organização (excepto para sistemas em tempo real). Este conceito de estrutura temporal permite detectar padrões e relações a longo prazo para auxiliar a tomada de decisão. O tempo é uma dimensão essencial que todos os DW devem suportar;

- Não volátil: Depois dos dados estarem inseridos no DW, os registos não podem ser modificados ou actualizados. Alguns dados considerados obsoletos pela organização podem ser eliminados (embora na generalidade nunca se remove dados de um DW) e as mudanças são registadas como sendo dados novos. Esta também é uma diferença importante em relação às BDO, que estão constantemente a realizar operações sobre os registos (inserções, alterações e remoções).

Normalmente, o DW é mantido separadamente da BDO da organização. Existe uma miríade de razões para fazer esta divisão. Uma delas reside na finalidade dos dois sistemas. Segundo Chaudhuri e Dayal (1997), os sistemas operacionais, ou *On-Line Transactional Processing* (OLTP), têm como finalidade centrarem-se no registo das transacções que ocorrem no seu funcionamento diário. Estas operações são estruturadas e repetitivas e consistem em transacções pequenas, atómicas e isoladas. Em oposição às BDO, segundo Han e Kamber (2006), um sistema de DW é normalmente apresentado como analítico, ou OLAP, porque é orientado a executivos ou analistas da informação que necessitam de efectuar análises e tomar decisões. O DW tem então como objectivo principal o armazenamento histórico e a integração dos dados provenientes das mais diversas fontes da organização, como visto anteriormente.

De forma a sintetizar as principais diferenças entre um DW e uma BDO, a Tabela 2.1, adaptada de Wu e Buchman (1997) mostra as suas características fundamentais.

	Base de dados Operacional	***Data Warehouse***
Função	• Operações diárias • OLTP	• Suporte à decisão • OLAP
***Design* da Base de Dados**	• Orientada às aplicações • Optimizada para actualizações	• Orientada aos assuntos • Optimizada para processamento de *queries*
Dados	• Correntes • Actualizados • Atómicos • Relacionais (Normalizados) • Isolados	• Históricos • Sumarizados • Multidimensionais • Integrados
Utilização	• Repetitivo • Dia-a-dia	• *Ad-hoc*
Acessos	• Leitura/Escrita • Transacções simples (envolvendo 1 a 3 tabelas)	• Maioria de Leitura • *Queries* complexas (envolvendo várias tabelas)

Tabela 2.1 - Base de dados Operacional vs *Data Warehouse*

É de realçar ainda que um dos principais problemas quando se constrói um DW prende-se com a distinção entre os dados operacionais e os dados informacionais (Gardner, 1998), isto porque a informação requerida para o suporte à decisão tem de estar relacionada com várias unidades de negócio, estando orientada para um determinado assunto da organização, contrariamente aos dados operacionais que são organizados de maneira a satisfazer os requisitos do processamento funcional (Santos & Ramos, 2009).

Depois de vistas as principais características de um DW, relembra-se que este é orientado à integração da informação relativa a determinados assuntos da organização e que portanto caracteriza a organização como um todo. Todavia, sempre que existe a necessidade de separar a informação, então armazena-se esta última num repositório de dados mais pequeno do que os DW, sendo designado de *Data Mart*.

Os *Data Marts* são então repositórios mais pequenos do que os DW que reúnem todos os dados de um subconjunto específico da organização, como por exemplo um departamento. Segundo Gardner (1998), essa diferenciação entre *Data Marts* e DW depende das necessidades da organização e decidir qual a arquitectura a implementar depende também do âmbito da informação para a tomada de decisão. Arriyachandra e Watson (2005) identificaram dez factores que podem afectar a decisão da selecção da arquitectura a implementar:

- A interdependência da informação entre as unidades organizacionais;
- A necessidade de informação por parte da gestão de topo;
- A urgência da necessidade de um DW;
- A natureza das tarefas dos utilizadores finais;
- As restrições de recursos;
- A visão estratégica do DW antes da sua implementação;
- A compatibilidade com os sistemas existentes;
- O *know-how* em Tecnologias de Informação (TI) dos recursos humanos da própria organização;
- As questões técnicas;
- Os factores político-sociais.

Os pontos de vista do *design* da arquitectura cabem normalmente em duas categorias: o *design* de um DW para toda a organização e o *design* de *Data Marts* (Golfarelli & Rizzi, 2009). Uma organização pode optar então pela implementação das três arquitecturas tradicionais e mais comuns apresentadas na Figura 2.2, Figura 2.3 e Figura 2.4.

Um DW organizacional é um DW de larga escala que é usado transversalmente por toda a organização para as tomadas de decisão. Esta característica faz com que os dados das várias fontes sejam integrados num formato *standard*. Por abranger toda a organização, um DW organizacional requer uma modelação detalhada no negócio, para este estar devidamente reflectido no repositório. Também pode levar anos a ser desenhado e construído.

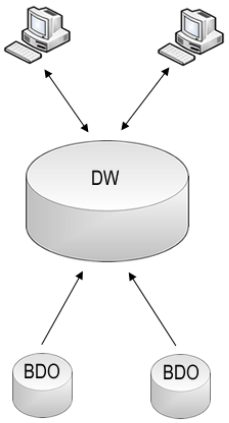


Figura 2.2 - *Data Warehouse* Organizacional

Ao contrário do DW organizacional, os *Data Marts* independentes não retratam todo o negócio da organização. Estes últimos integram dados de diversas fontes, como acontece com os DW organizacionais, mas apenas alguns subconjuntos de dados que são relevantes para um grupo específico de utilizadores (e.g. departamentos). O custo elevado dos DW organizacionais faz com que muitas organizações utilizem esta versão *low-cost* de DW cujo ciclo de implementação é significativamente mais reduzido.

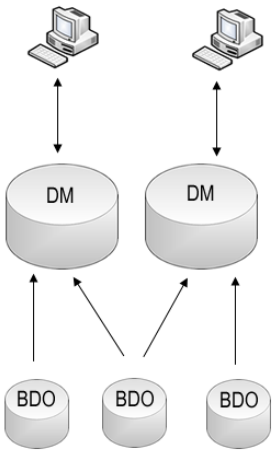


Figura 2.3 - *Data Marts* Independentes

Os *Data Marts* dependentes, ao invés dos *Data Marts* independentes, são alimentados directamente a partir do DW organizacional. Têm então como vantagem utilizar um modelo de dados consistente e fornece dados com qualidade. Os *Data Marts* dependentes suportam a

ideia de existir um modelo de dados transversal à organização que tem de ser construído antes da implementação destes primeiros. Esta arquitectura assegura também aos utilizadores que estes visualizem a mesma versão dos dados que é acedida por todos os outros utilizadores do DW organizacional, apesar de poder existir uma ligeira latência temporal (que pode ser resolvida com a actualização do *Data Mart*).

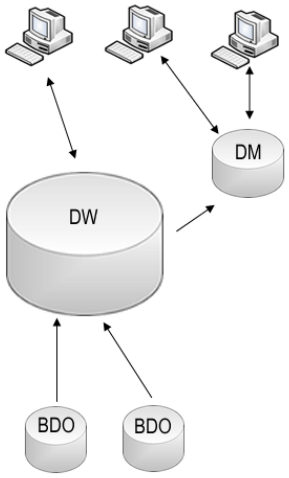


Figura 2.4 - *Data Marts* Dependentes

Para além destas três arquitecturas mais tradicionais, Arriyachandra e Watson (2006) referem também outro tipo de arquitectura alternativa: a *Federated Architecture*, apresentada na Figura 2.5. Utiliza todos os meios possíveis para integrar recursos analíticos de várias fontes para atender às necessidades de mudança ou condições de negócios da organização. Essencialmente, a *Federated Architecture* envolve a integração de sistemas díspares. As estruturas de apoio à decisão existentes são mantidas e os dados são acedidos dessas fontes quando necessário. Esta arquitectura é apoiada essencialmente por fabricantes de *middleware* que fornecem capacidades de *queries* distribuídas e de *joins*. Essas ferramentas baseadas em XML (*eXtensible Markup Language*) oferecem ao utilizador uma visão global dos dados, incluindo os DW, os *Data Marts*, os *Websites*, os documentos e as BDO (Turban, et al., 2010). Devido a problemas de desempenho e qualidade dos dados, a maioria dos especialistas, segundo Eckerson (2005), concorda que a *Federated Architecture* funciona como suplemento às arquitecturas mais usuais, mas não como substituta.

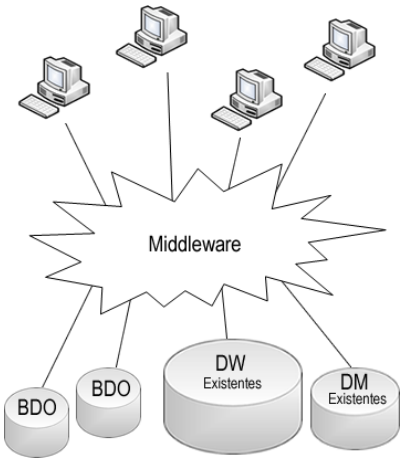


Figura 2.5 - *Federated Architecture*

Como complemento aos esclarecimentos sobre as várias arquitecturas possíveis, a Tabela 2.2, adaptada de Eckerson (2003a), expõe algumas das forças e fraquezas das quatro abordagens apresentadas anteriormente.

	Vantagens	**Desvantagens**
***Data Marts* independentes**	• Fácil de construir do ponto de vista da organização • Fácil de construir do ponto de vista técnico	• Vista organizacional indisponível • Custos com dados redundantes • Custos elevados com o processo de ETL • Custos elevados com aplicações • Custos com o DBA[2] e custos operacionais elevados
***Data Marts* dependentes**	• Permite uma mais fácil parametrização das interfaces e dos relatórios	• Vista organizacional muito difícil • Custos com dados redundantes • Custos com o DBA e custos operacionais elevados • Latência de datas
***Data Warehouse* organizacional**	• Vista organizacional • *Design* consistente e qualidade dos dados • Reutilização de dados	• Requer uma forte liderança e visão
Federated Architecture	• Não há necessidade de processo de ETL • Não há necessidade de separar plataformas	• Apenas viável para um volume de dados não muito grande • Problemas com *Metadatas* • Problemas com a largura de banda da rede e a complexidade dos *joins*

Tabela 2.2 - Vantagens e desvantagens de algumas arquitecturas de *Data Warehousing*

[2] DBA: *Data Base Administrator* – Administrador de Base de Dados

Os *Metadatas* são uma parte importante de qualquer sistema de *Data Warehousing*. São dados sobre os dados (Sen, 2004). Uma vez que o DW reflecte o modelo de negócio de uma organização, é necessária a gestão dos seus vários tipos de *Metadatas*. Muitas vezes, um repositório próprio para os *Metadatas* é usado para armazenar e gerir todos os *Metadatas* associados ao DW. Este repositório permite a partilha dos *Metadatas* sobre ferramentas e processos de *design*, concepção, utilização e administração de um DW. Pode incluir informações como: a descrição das fontes de dados; as ferramentas de *front-end* e *back-end*; a definição do esquema do DW; as dimensões e hierarquias; *queries* e relatórios predefinidos; a localização e os conteúdos dos *Data Marts*; as regras de extracção, limpeza e carregamento e refrescamento dos dados; os perfis de utilizadores e políticas de controlo de acesso e autorizações.

Como referido anteriormente aquando da comparação entre as BDO e os DW, estes últimos usam uma modelação multidimensional como estrutura. Esta modelação multidimensional proporciona uma estrutura de base de dados fácil de utilizar e de compreender e optimiza o processamento de *queries*, ao invés do que acontece com o modelo relacional de base de dados que optimiza o processamento de actualizações. Esta modelação multidimensional pode ser conseguida através da implementação de três esquemas (Santos & Ramos, 2009):

- Esquema em Estrela;
- Esquema em Constelação;
- Esquema em Floco de Neve.

A forma mais comum e mais simples de modelar um DW é através do esquema em estrela, cuja representação pode ser vista na Figura 2.6.

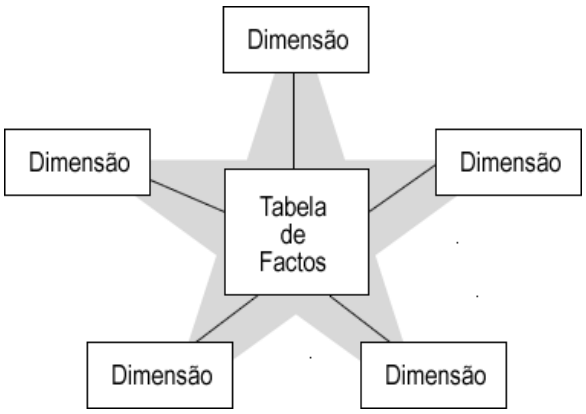


Figura 2.6 - Esquema em Estrela

Este esquema integra uma tabela de factos, que pode ser considerada o centro da estrela, e várias tabelas de dimensões (um esquema em estrela pode ter um número qualquer de dimensões) que podem ser vistas como as pontas da estrela. As tabelas de dimensões estão todas ligadas à tabela de factos com ligações do tipo chave primária / chave estrangeira, mas não estão ligadas entre elas. As tabelas dimensões são geralmente tabelas não normalizadas e permitem analisar os dados sobre diversas perspectivas. Esta última característica provém do facto de ser um modelo multidimensional e deste poder ser visualizado como um cubo com várias dimensões, permitindo a navegação entre categorias específicas, *drill-downs*, entre outras técnicas. Este conceito de cubo multidimensional é apresentado mais detalhadamente na próxima subsecção aquando da exposição da tecnologia OLAP.

A tabela de factos corresponde geralmente à componente do negócio que se pretende analisar (e.g. vendas, compras) e integra dois tipos de atributos:

- Factos: São atributos numéricos que podem ser analisados usando várias funções estatísticas. Existem três tipos de factos, estes podem ser aditivos, semi-aditivos e não-aditivos. Os factos aditivos são os que podem ser somados relativamente a todas as dimensões de um esquema em estrela. Os factos semi-aditivos apenas podem ser somados relativamente a algumas dimensões (ou mesmo a apenas uma única dimensão). Os factos não-aditivos são os que não podem ser somados de acordo com qualquer uma das dimensões;
- Chaves Estrangeiras: Permitem a ligação entre as várias dimensões existentes e, consequentemente, a visão multidimensional.

É de realçar ainda que a tabela de factos representa quase a totalidade do espaço ocupado pelo DW, contendo muitos mais registos mas muito menos atributos do que as dimensões. Na Tabela 2.3 são sintetizadas as principais características da tabela de factos e das tabelas de dimensões.

Tabela de Factos (TF)	Tabelas de Dimensões (TD)
• Normalizada • Mais registos do que as TD • Contém poucos atributos • Integra Factos • Integra Chaves Estrangeiras	• Não normalizada • Menos registos do que a TF • Muitos atributos • Número de dimensões igual ao número de perspectivas sobre as quais se pretende analisar os dados

Tabela 2.3 - Tabela de Factos vs Tabelas de Dimensões

Um esquema em Constelação é muito parecido com o esquema em Estrela. A única diferença é que este integra várias tabelas de factos que partilham uma ou diversas dimensões. Este esquema pode ser visto como um conjunto de esquemas em estrelas, como na analogia com a astronomia (constelação – conjunto de estrelas), que se ligam através de dimensões comuns. Uma representação deste esquema pode ser vista na Figura 2.7.

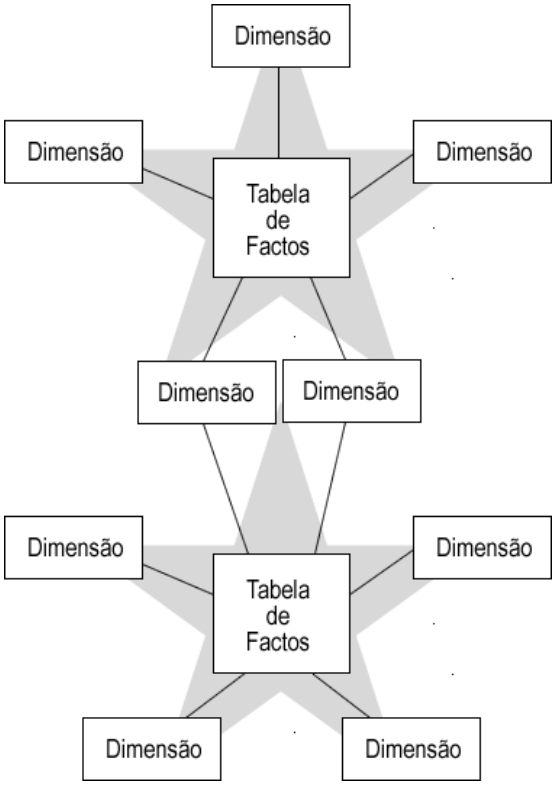


Figura 2.7 - Esquema em Constelação

Por sua vez, o esquema em Floco de Neve é um esquema em Estrela que contém algumas (ou todas) das suas dimensões normalizadas. Isto significa que este deixa de ter uma estrutura regular, isto porque as suas dimensões podem ter um número diferente de hierarquias (ou sub-dimensões). Segundo Moody e Kortink (2003), este esquema tem como vantagem indicar explicitamente a estrutura das suas dimensões, o que não acontece com o esquema em Estrela ou em Constelação que integram um conjunto não estruturado de dados (não normalizado). Contudo, como o esquema em Floco de Neve se encontra normalizado, as maiores desvantagens do mesmo centram-se na dificuldade de interpretação do esquema e à perda de desempenho no processamento de *queries* (Han & Kamber, 2001). Assim, este esquema não é tão popular como o esquema em Estrela. Uma representação deste esquema pode ser vista na Figura 2.8.

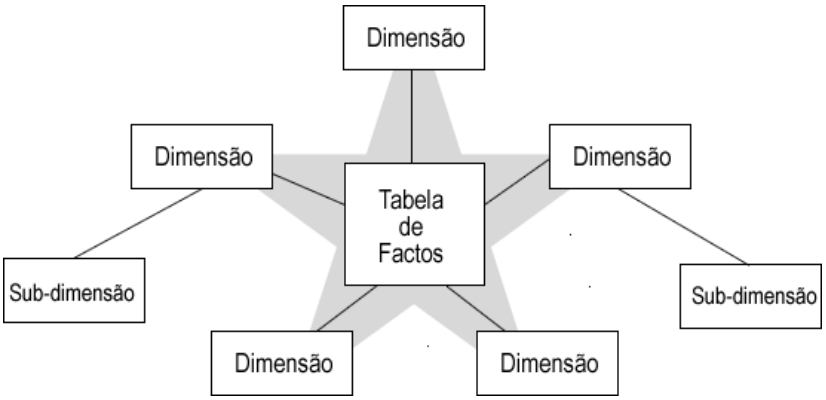


Figura 2.8 - Esquema em Floco de Neve

Depois de vistos e explicitados os esquemas multidimensionais possíveis para a construção de um DW, de seguida vai ser abordada a maneira dos dados passarem das BDO para os DW.

Segundo Vassiliadis et al. (2002), as ferramentas de ETL são ferramentas especializadas em lidar com a homogeneização dos dados, da sua limpeza e do seu posterior carregamento para o DW. Durante a primeira fase, os dados são extraídos das fontes internas e externas disponíveis. Pode aqui ser feita uma distinção lógica entre a primeira extracção, onde todos os dados disponíveis relativos a todos os períodos passados são carregados para um DW vazio, e as subsequentes extracções incrementais que actualizam o DW utilizando novos dados que se tornam disponíveis ao longo do tempo (refrescamento). Esta selecção dos dados a serem importados para o DW é baseada no *design* do DW, que por sua vez depende da informação necessária para as análises do sistema de BI num domínio de aplicação específico.

O objectivo da fase de transformação e limpeza dos dados é melhorar a qualidade dos dados extraídos das diferentes fontes, através de correcções de inconsistências, erros e valores em falta. Algumas das maiores deficiências nos dados que são removidas nesta fase do processo de ETL são as seguintes:

- Inconsistência entre os valores de diferentes atributos que têm o mesmo significado;
- Duplicação de dados;
- Dados em falta;
- Existência de valores errados.

Durante a fase de limpeza dos dados são também aplicadas regras automáticas pré-definidas para corrigir os erros mais recorrentes. Em muitos casos, dicionários com termos válidos são usados para substituir os termos incorrectos com base no nível de similaridade (Vercellis, 2009). Para além disso, durante a fase de transformação, conversões adicionais ocorrem para garantir a homogeneidade e integração entre os dados provenientes das várias fontes de dados.

Finalmente, depois dos dados serem extraídos e transformados, estes são carregados para as tabelas do DW e torna-os disponíveis para os analistas e as aplicações de apoio à decisão.

A Figura 2.9 representa a arquitectura de um sistema de ETL apresentada por Vassiliadis et al. (2002).

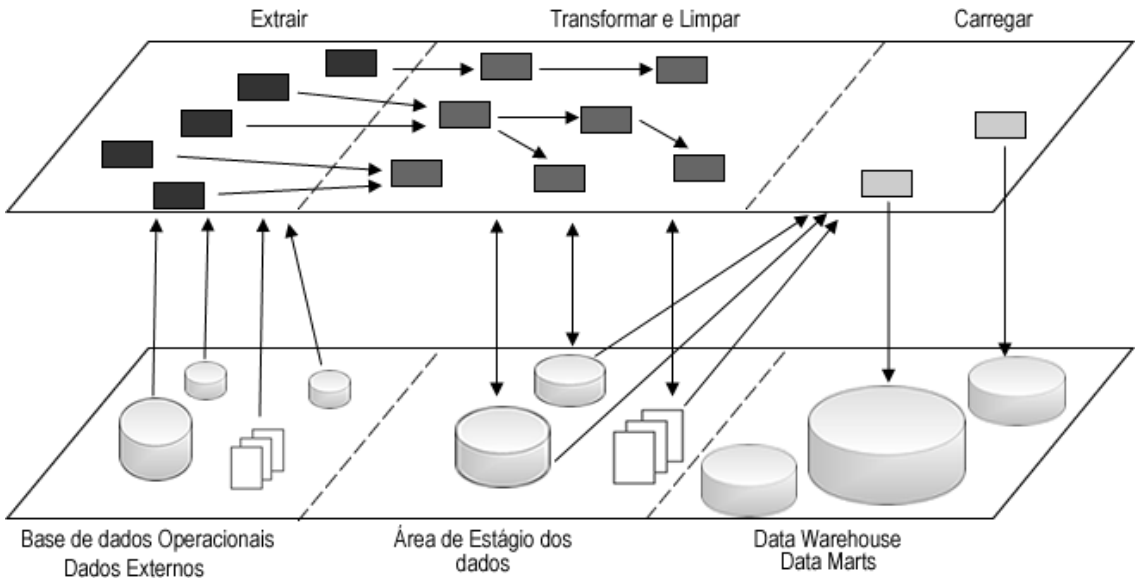


Figura 2.9 - O processo de ETL

No nível inferior encontram-se os vários repositórios pelos quais os dados passam durante todo esse processo, enquanto que no nível superior estão representadas as operações efectuadas sobre os dados. É de realçar então que os dados passam por repositórios intermédios e são armazenados temporariamente na área de estágio dos dados (*Data Staging Area* - DSA) durante o processo de transformação e limpeza. Depois de concluído este processo, os dados presentes na DSA são carregados para o DW.

Depois de vistos os conceitos mais importantes relacionados com o *Data Warehousing*, de seguida é brevemente apresentada a metodologia mais conhecida para a implementação

de sistemas deste tipo: a metodologia Kimball. A Figura 2.10, adaptada de Kimball et al. (1998) retrata todo o processo desta metodologia. As actividades que compõem este processo são de seguida muito resumidamente descritas.

Na primeira actividade da metodologia, é realizado o plano do projecto (riscos associados às diferentes etapas, motivação da organização, critérios de sucesso, retorno do investimento do DW, etc.). A actividade seguinte está relacionada com a identificação de todo o tipo de requisitos iniciais para o sistema de DW. A Definição dos Requisitos de Negócios serve como ponto de partida para três actividades que decorrerão em paralelo. Na primeira, são definidos os critérios para a selecção de produtos (hardware, base de dados, sistemas de carregamento de dados, ferramentas de acesso aos dados) que implementam essa arquitectura. No seguimento desta actividade os produtos são seleccionados e instalados. Na segunda, ocorre a modelação dimensional. Daí resultam os esquemas em estrela, e as suas variações, das bases de dados. Seguem-se o modelo físico e a concepção e desenvolvimento das DSA assim como todo o processo de carregamento e refrescamento dos dados. Na terceira, são arquitectadas e construídas as aplicações que serão utilizadas para aceder ao sistema de DW. Na actividade de Implementação, tem que ser garantida a construção e a população do sistema de DW, assim como as suas aplicações de acesso. É também necessário definir a estratégia de formação e de suporte para os utilizadores finais. A actividade seguinte tem como propósito monitorizar e gerir o sistema de DW, assim como definir prioridades em termos de alterações que passarão de novo por este ciclo. De salientar ainda que logo que se dá início ao projecto é despoletada uma actividade que vai gerir o mesmo durante todo este ciclo.

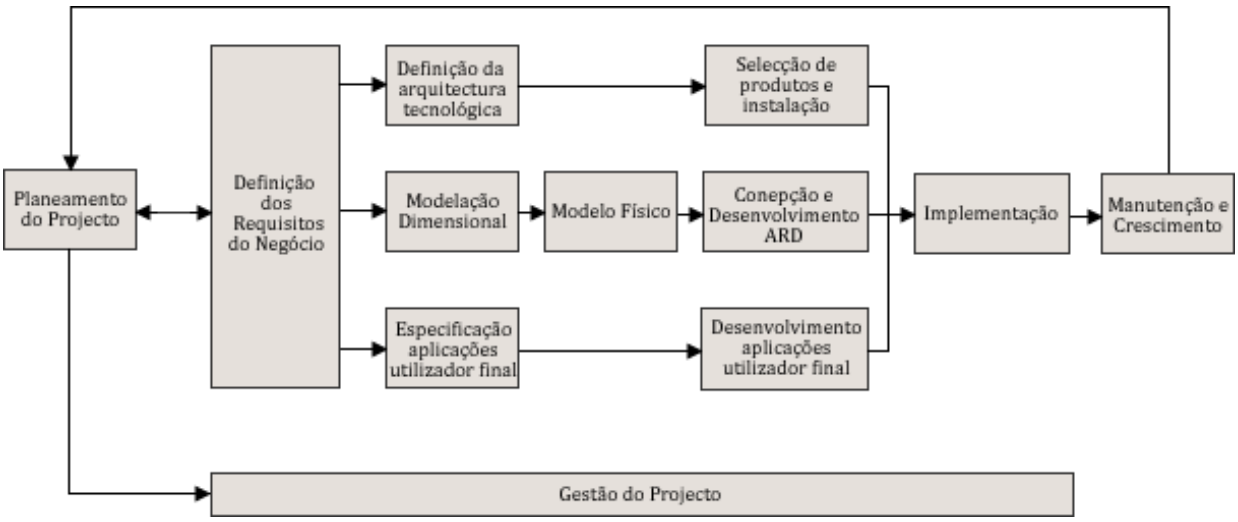


Figura 2.10 - Metodologia Kimball

2.1.2.2 Os Sistemas de Processamento Analítico de Dados

Os sistemas de processamento analítico de dados são uma das várias tecnologias que podem ser usadas para a exploração dos dados e da informação de um DW, sendo juntamente com o DM, cujos conceitos associados são apresentados na próxima subsecção, a tecnologia mais comummente utilizada.

A tecnologia OLAP permite realizar análises multidimensionais aos dados e oferece a capacidade de realizar cálculos complexos, analisar tendências e de modelar os dados refinadamente. Graças a estas características, o OLAP tornou-se rapidamente numa tecnologia fundamental para qualquer sistema de BI e qualquer SAD orientado aos dados. O OLAP permite que os utilizadores finais façam análises *ad-hoc* em dados com múltiplas dimensões, fornecendo-lhes assim a informação e o conhecimento que necessitam para uma melhor tomada de decisão.

A modelação multidimensional faz com que essa tecnologia crie cubos para analisar a informação necessária à tomada de decisão sobre várias perspectivas. Como referido na subsecção anterior, aquando da apresentação da modelação multidimensional, os factos (presentes na tabela de factos) podem ser analisados pelas várias dimensões definidas. Quando uma tabela de factos está associada a *n* dimensões, sendo *n* maior que 3, o cubo de dados correspondente constitui uma estrutura em *n* dimensões que não pode ser representada graficamente. Todavia, segundo Vercellis (2009), num cubo com quatro dimensões, por exemplo, é possível obter quatro vistas lógicas compostas por cubos de três dimensões, denominados de cubóides, dentro do cubo de quatro dimensões, fixando os valores de uma dimensão. Geralmente, é possível obter uma grelha de cubóides, cada um deles correspondente a um nível diferente de consolidação ao longo das várias dimensões.

A Figura 2.11 mostra a representação de um cubo no qual um dos factos, as vendas por exemplo, é analisado sob três dimensões (Produto, Tempo e Região). A Figura 2.12, adaptada de Vercellis (2009), ilustra um exemplo de uma grelha de cubóides obtidos por um cubo de quatro dimensões.

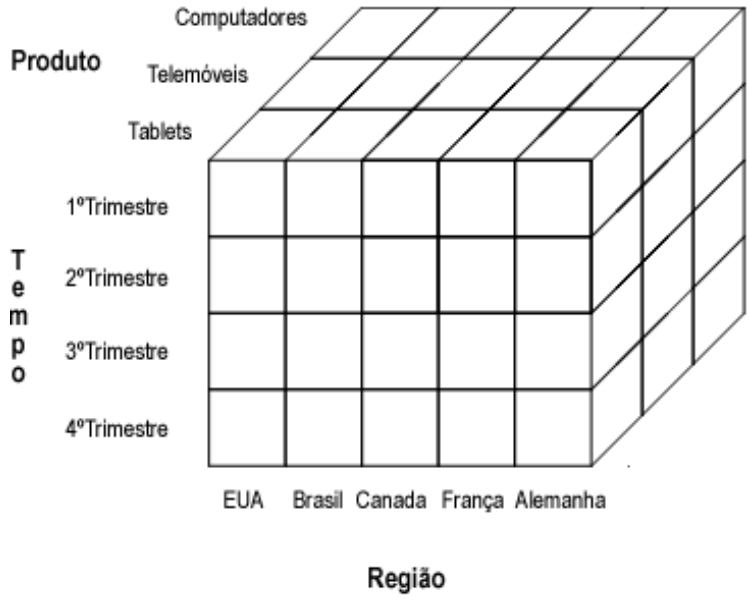


Figura 2.11 - Exemplo de um cubo de três dimensões

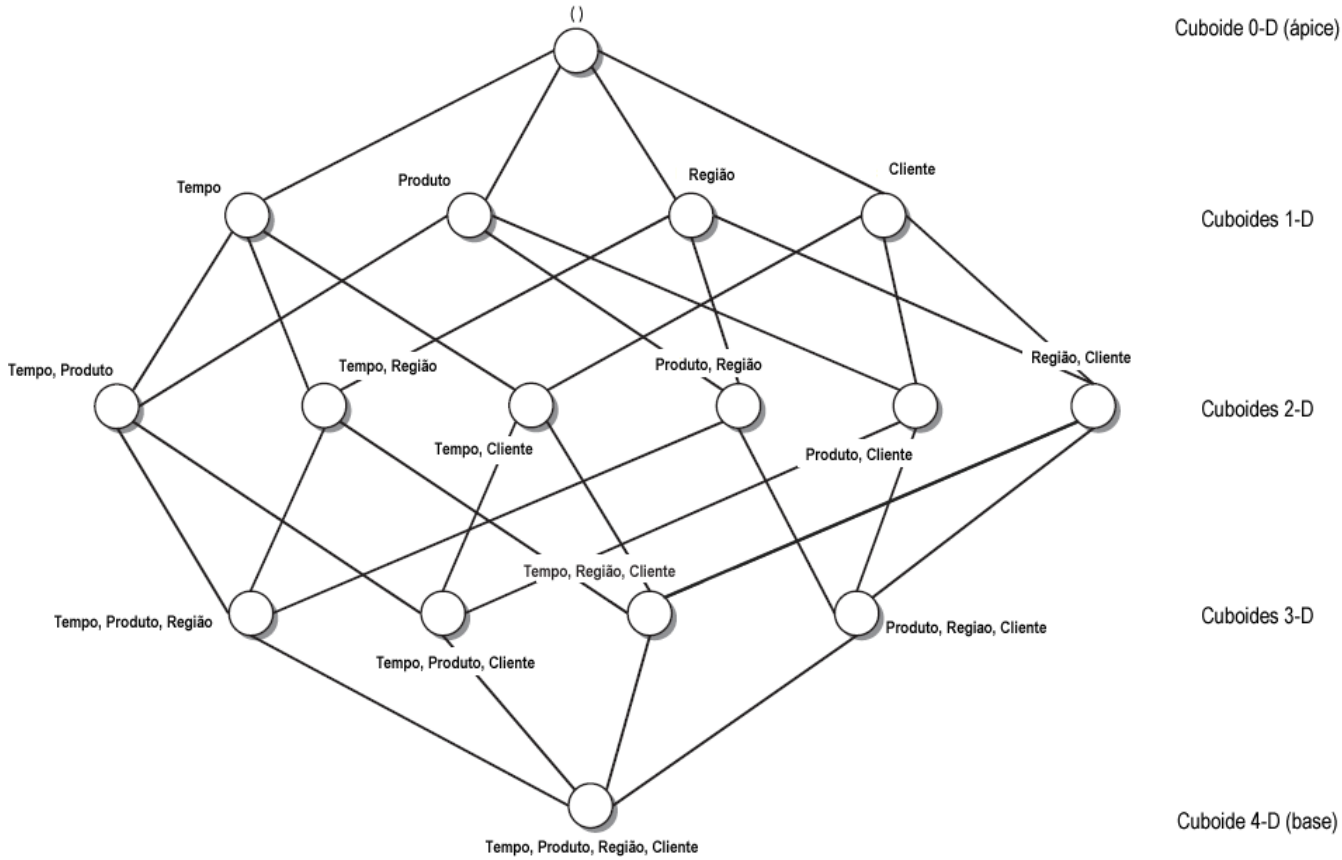


Figura 2.12 - Grelha de cubóides derivada de um cubo de quatro dimensões

Existem três tipos de servidores OLAP que permitem analisar multidimensionalmente os dados a partir de qualquer tipo de repositório (bases de dados relacionais, DW e *Data Marts*, etc.) Os servidores OLAP podem ser (Han & Kamber, 2001):

- MOLAP (*Multidimensional* OLAP): Estes servidores suportam vistas multidimensionais de dados que já estão armazenados sob forma de cubos multidimensionais. Uma das vantagens de usar servidores MOLAP é o seu excelente desempenho. Estes são construídos com o intuito de permitir uma rápida indexação a dados pré-processados;
- ROLAP (*Relational* OLAP): São servidores que servem de intermédios entre uma base de dados relacional e as ferramentas de *front-end* das análises de dados. Este tipo de servidor consegue lidar com grandes quantidades de dados. A limitação do tamanho de dados da tecnologia ROLAP é a própria limitação da base de dados relacional subjacente, ou seja, o próprio ROLAP não coloca nenhuma limitação na quantidade de dados. Por outro lado, o seu desempenho pode ser muito mais lento do que o do servidor MOLAP, isto porque cada *report* é essencialmente uma consulta SQL (ou várias consultas SQL) à base de dados relacional e portanto o tempo de consulta pode ser demorado se o tamanho da base de dados relacional subjacente for grande;
- HOLAP (*Hybrid* OLAP): Este servidor híbrido combina as tecnologias MOLAP e ROLAP, beneficiando da grande escalabilidade do ROLAP e da maior velocidade de computação do MOLAP. Isto significa que um servidor HOLAP pode armazenar grandes quantidades de dados numa base de dados relacional, enquanto que as agregações são mantidas separadamente num repositório MOLAP.

Geralmente, as análises recorrendo a ferramentas de processamento analítico de dados são baseadas em hierarquias de conceitos para consolidar os dados e para criar vistas lógicas ao longo das dimensões de um DW. Por exemplo uma dimensão “local” pode originar uma hierarquia perfeitamente ordenada (país, distrito, concelho, freguesia). Estas hierarquias de conceitos são usadas para executar operações de visualização sobre cubos de um DW. Existem diferentes tipos de operações que podem ser executadas sobre os cubos. As mais comuns são as seguintes (Han & Kamber, 2001):

- *Roll-Up*: A operação de *roll-up* (também conhecida como *drill-up*) consiste numa agregação de dados do cubo que pode ser obtida de duas maneiras. (1) Subindo para um nível mais elevado da hierarquia de uma dada dimensão. Por exemplo, para a dimensão "localização" é possível subir do nível "País" para o nível "Região". (2) Eliminando uma dimensão à analise do cubo;
- *Drill-Down*: Esta operação, também referida como *roll-down*, é a operação oposta do *roll-up*. Permite navegar através do cubo a partir de dados agregados para dados mais detalhados, possibilitando assim uma visão mais pormenorizada dos dados. Sendo a operação inversa do *roll-up*, esta operação também pode ser realizada de duas maneiras. (1) Descendo um nível na hierarquia de uma dada dimensão. Por exemplo, para a dimensão "Tempo" é possível mudar do nível "Trimestre" para o nível "Mês". (2) Adicionando uma dimensão ao cubo;
- *Slice and Dice*: Através da operação *slice*, o valor de um atributo de uma dimensão é seleccionado e fixado, sendo analisado em relação às restantes dimensões. O *dice* permite definir um sub-cubo seleccionando atributos em duas ou mais dimensões;
- *Pivot*: A operação de *pivot*, também chamada de *rotate*, permite rodar os eixos dos dados de forma a disponibilizar uma representação alternativa dos mesmos.

A Figura 2.13 e a Figura 2.14 ilustram as operações descritas anteriormente.

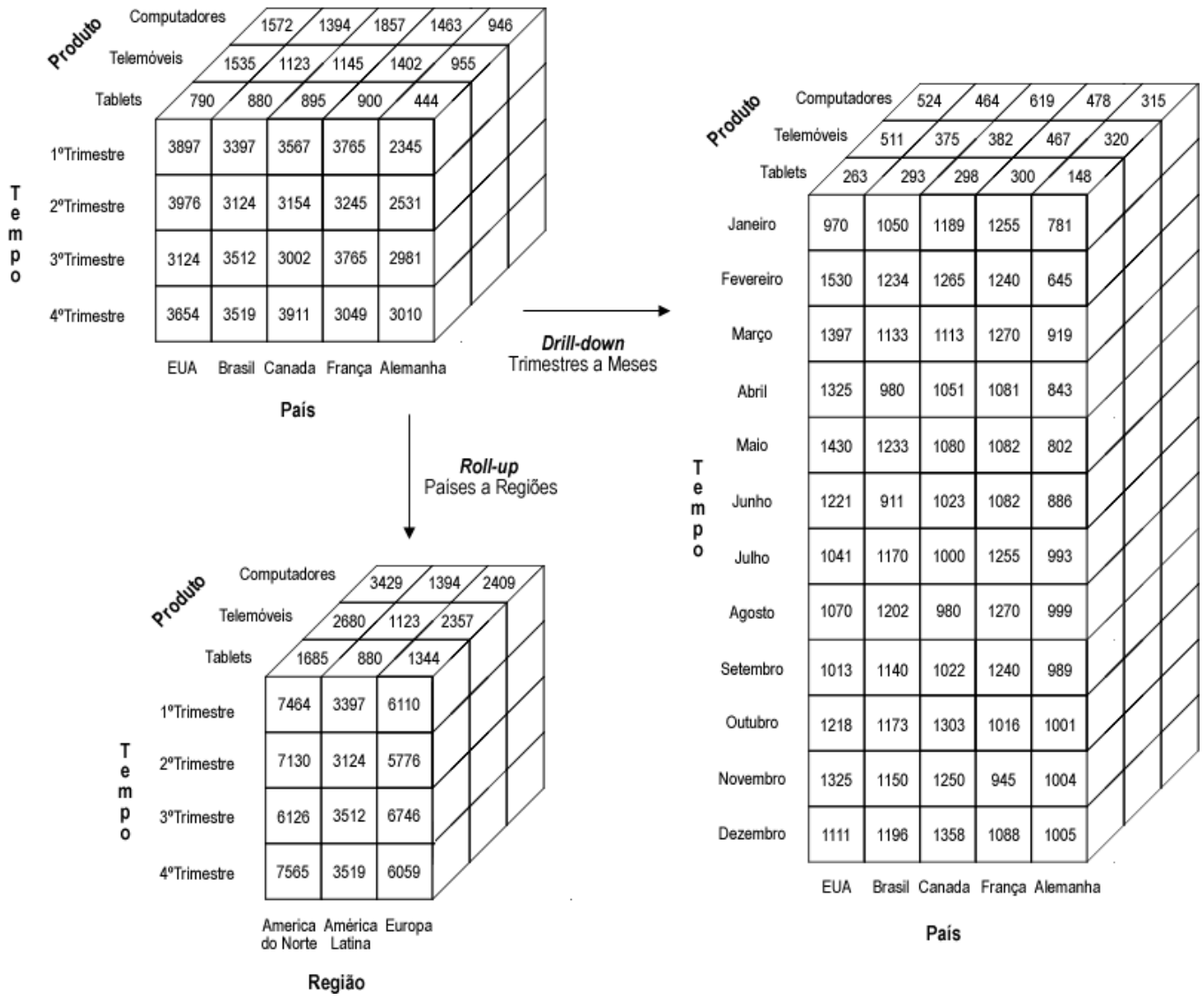


Figura 2.13 - Manipulação de cubos: *Roll-up* e *Drill-down*

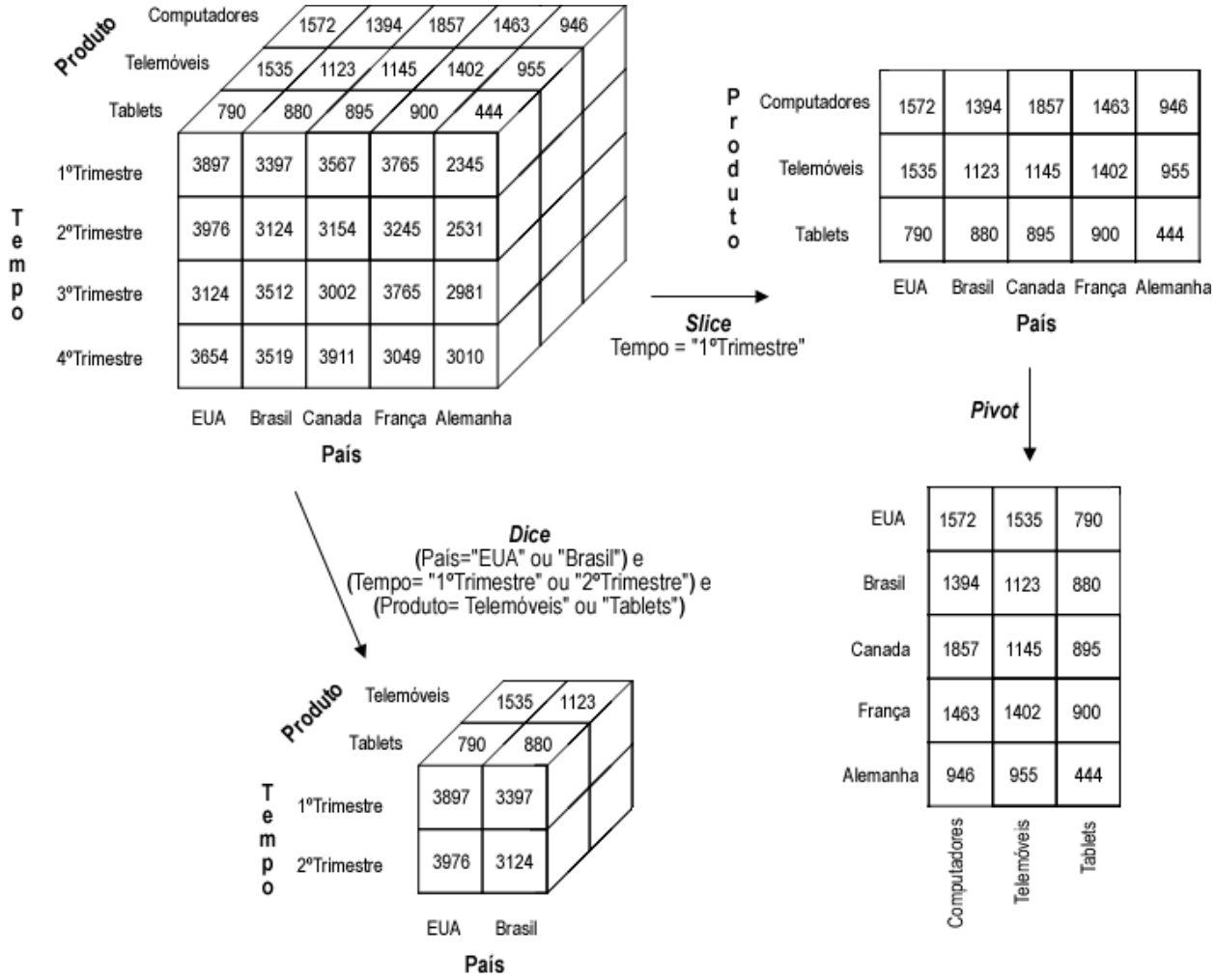


Figura 2.14 - Manipulação de cubos: *Slice*, *Dice* e *Pivot*

2.1.2.3 *Data Mining*

Depois de vistos os conceitos subjacentes ao *Data Warehousing* e ao OLAP, nesta subsecção serão apresentados os conceitos mais importantes relacionados com o DM.

A principal diferença entre o DM e as outras ferramentas de análise de dados, como por exemplo as análises multidimensional, ou OLAP, cujo conceito foi explicado na subsecção anterior, está na forma como estas exploram os dados. Enquanto que nas análises multidimensionais o utilizador tem de construir hipóteses sobre relações entre os dados e tem de confiar na sua intuição e habilidade em elaborar essas hipóteses (manipulando os atributos e refinando a análise baseado nos resultados das consultas à base de dados), no DM o próprio processo pode ficar responsável pela geração de hipóteses, garantido maior rapidez,

aperfeiçoamento e fiabilidade aos resultados. O DM, através do uso de algoritmos específicos ou de mecanismos de pesquisa, tenta descobrir padrões discerníveis e tendências nos dados e inferir regras para os mesmos (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurasamy, 1996). Essa análise pode permitir um acréscimo de conhecimento sobre o negócio, ao ir além da simples análise aos dados guardados explicitamente.

O termo DM pode ser definindo de várias maneiras. Segundo Fayyad et al. (1996), o DM é o processo de encontrar padrões e relações em bases de dados de grandes dimensões, previamente desconhecidos e potencialmente interessantes. A procura desses padrões de interesse é feita numa determinada forma de representação, ou conjunto de representações (classificação, árvores de decisão, regras de indução, regressão, segmentação, etc.). Os mesmos autores também definem o DM como sendo a aplicação de algoritmos para a extracção de padrões dos dados sem os passos adicionais do processo Descoberta de Conhecimento em Base de Dados (DCBD), como a inclusão de conhecimento anterior e a interpretação correcta dos resultados. Chang et al. (2001) definem ainda o DM como sendo o processo de extrair informação ou conhecimento de conjuntos de dados para auxiliar a tomada de decisão.

Existem várias metodologias que são utilizadas no processo de DCBD, das quais, devido ao âmbito deste projecto de dissertação, vale a pena realçar a mais conhecida: o CRISP-DM (*CRoss-Industry Standard Process for Data Mining*). A Figura 2.15, adaptada de Chapman et al. (2000) retrata todo o processo desta metodologia que permite o retrocesso a etapas anteriores para incluir novos dados ou alterar decisões, dependendo essencialmente do resultado e do desempenho das outras fases. Seguidamente apresentam-se muito resumidamente as seis fases definidas por esta metodologia.

- Compreensão do negócio: Nesta fase inicial pretende-se compreender os objectivos e os requisitos do projecto do ponto de vista do negócio. Esta compreensão levará à conversão destes objectivos do negócio em objectivos de DM. Deve ser elaborado um plano para atingir os mesmos;
- Compreensão dos dados: Esta fase começa com a recolha de dados e prossegue com tarefas de exploração com o intuito de se compreender os mesmos, de identificar problemas na qualidade dos dados ou de detectar subconjuntos relevantes que serão posteriormente analisados com vista à extracção de conhecimento implícito;

- Preparação dos dados: Esta fase cobre todas as actividades para a construção do conjunto de dados que será analisado pelos algoritmos de DM. As tarefas para a preparação dos dados incluem a selecção de tabelas, registos e atributos, assim como a transformação e limpeza dos dados para a sua análise posterior;
- Modelação: É nesta fase que as várias técnicas de modelação são seleccionadas, aplicadas e os seus parâmetros ajustados no sentido de melhorar os resultados obtidos. Como algumas técnicas têm requisitos específicos, como a análise de determinado tipo de dados, pode ser necessário retroceder à fase de preparação dos dados para convertê-los num determinado formato;
- Avaliação: Esta fase tem como intuito avaliar a utilidade dos modelos obtidos na fase anterior, rever os passos executados nas suas construções e verificar se permitem atingir os objectivos do negócio, ou se alguma questão importante para o negócio não tenha sido considerada;
- Implementação: A criação do modelo não significa o fim do projecto. O conhecimento extraído tem de ser organizado e apresentado para que possa ser utilizado. Dependendo dos requisitos, esta fase pode ser tão simples como a geração de um relatório ou pode ser tão complexa como implementar todo o processo de DM.

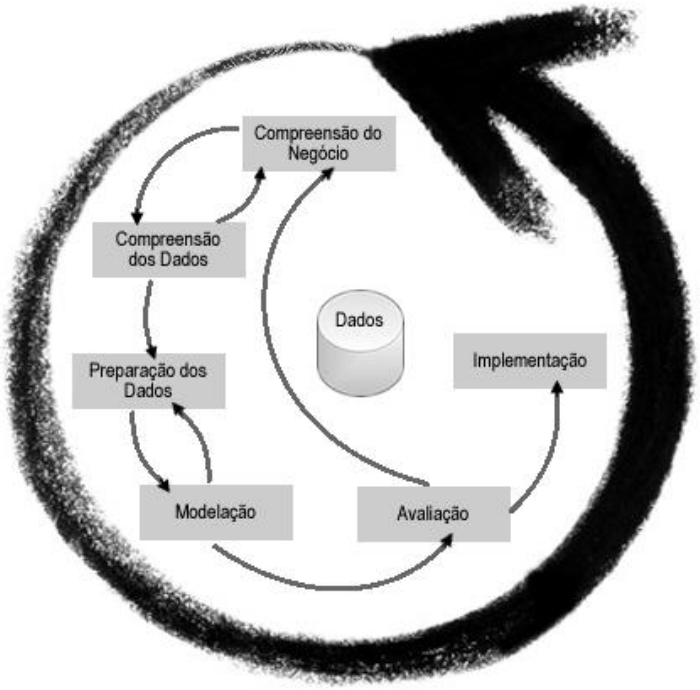


Figura 2.15 - Metodologia CRISP-DM

O DM não é uma disciplina nova, pode ser considerada como uma nova definição para o uso de várias disciplinas em conjunto (Turban, et al., 2010), sendo a intersecção de áreas como a estatística e matemática, inteligência artificial, base de dados e DW, reconhecimento de padrões. Os modelos e técnicas utilizadas no DM provêm dessas áreas, tendo como base a extracção de conhecimento ou padrões de informação dos dados em grandes bases de dados.

Existem dois tipos de abordagens ou estratégias que podem ser usadas para fornecer informação relevante em DM (Berry & Linoff, 2004):

- Abordagem directa, *top-down* ou modelo de verificação (quando se sabe o que pesquisar);
- Abordagem indirecta, *bottom-up* ou modelo de descoberta (quando não se sabe qual o objectivo da pesquisa).

A primeira abordagem é usada quando o utilizador formula uma hipótese e testa a sua validade com os dados, para afirmar ou negar essa hipótese. Não é criada informação nova no processo de pesquisa, mas as consultas aos dados retornam registos para afirmar ou negar a hipótese. A procura torna-se iterativa, pois como a ênfase está no utilizador, este pode reformular novas hipóteses para esta ser refinada. Na abordagem indirecta, é o sistema que descobre automaticamente a informação importante que está implícita e oculta nos dados. Contrariamente à primeira abordagem, esta não tem como ênfase o utilizador, isto porque os dados são pesquisados de maneira a encontrar tendências e padrões sem a orientação de um utilizador. Esta é a abordagem mais comum em DM.

Depois de vistas as estratégias que podem ser usadas, vão ser apresentados os vários objectivos (ou tarefas) do DM. À semelhança das duas abordagens que podem ser seguidas em DM, as suas tarefas associadas enquadram-se também em duas categorias (Berry & Linoff, 2000), dependendo essencialmente do tipo de objectivo da tarefa de DM:

- Descrição: Pode permitir aumentar o conhecimento acerca dos dados analisados, identificando regras que os caracteriza;
- Previsão: Identifica modelos capazes de prever o valor de uma variável, utilizando para isso determinados atributos das bases de dados.

Santos e Ramos (2009) acreditam que o melhor modelo de previsão é aquele que apresenta a precisão mais elevada, isto é, permite uma percentagem de acerto superior aos

outros modelos, mesmo que estes últimos tenham sido mais fáceis de perceber e de obter. Por sua vez, o melhor modelo de decisão será aquele que permite adquirir um conhecimento mais alargado dos dados analisados, e não obrigatoriamente aquele que obtém resultados mais precisos em relação à confiança do modelo.

A Classificação, a Regressão Linear, a Segmentação (ou *Clustering*), a Associação (ou Dependência), a Sumariação e a Detecção de Desvios constituem as seis principais tarefas de DM (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

A Classificação corresponde a encontrar uma função que associe um caso a uma classe dentro de diversas classes discretas (números inteiros, conjunto finito) de classificação, de forma a catalogar um novo objecto de acordo com um modelo de classificação (Thuraisingham, 1999). Esta tarefa de DM permite então enquadrar um conjunto de dados dentro de classes predefinidas, identificando a classe a que cada um pertence. As técnicas usadas na Classificação utilizam conjuntos de treino com exemplos pré-classificados com o objectivo de construir modelos adequados à descrição das classes que posteriormente são aplicados em dados não classificados (conjuntos de teste). Estas classes representam então o conjunto de valores possíveis, explícitos nos dados analisados, e o atributo corresponde ao valor de saída do processo de Classificação. Como o atributo e as classes do processo de Classificação são conhecidas desde o início, esta tarefa é considerada uma tarefa de aprendizagem supervisionada.

Para exemplificar este conceito, é utilizado um exemplo de classificação para determinar se os empréstimos bancários são concedidos ou não (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). O exemplo é baseado em 23 casos de pedidos de empréstimo e como atributos são considerados o valor do empréstimo e o rendimento do indivíduo. Os dados foram classificados em duas classes: os maus pagadores e os bons pagadores, representados por X e O, respectivamente, na Figura 2.16. Portanto, dividindo os dados através de uma simples classificação linear, obtêm-se duas classes. Assim, o banco poderá decidir a atribuição dos futuros empréstimos.

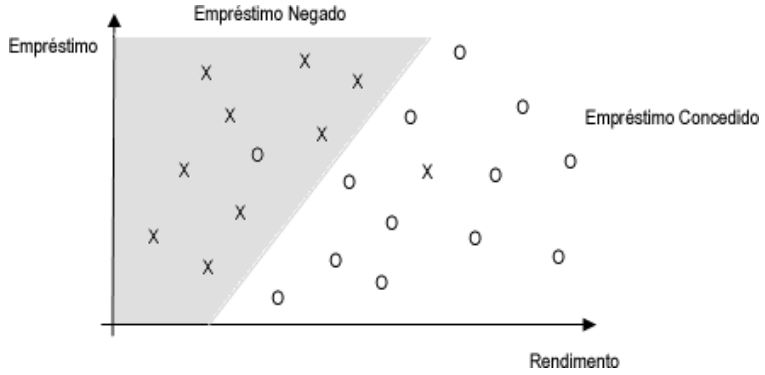


Figura 2.16 - Exemplo de Classificação

A Classificação é um dos objectivos ou tarefas de DM mais comum, sendo as Árvores de Decisão (AD), a Indução de regras (IR), os Algoritmos Genéticos e as Redes Neuronais Artificiais (RNA) as técnicas mais aplicadas (Chapman, et al., 2000).

Ao contrário da Classificação, que é usada para prever valores discretos, a Regressão Linear é utilizada para prever atributos com valores contínuos. Trata-se então de encontrar uma função para uma previsão de uma variável, ou seja, consiste na procura de uma função que represente comportamentos de variáveis de uma forma aproximada (Santos e Azevedo, 2005). A Regressão Linear permite a discriminação dos dados através da combinação dos atributos de entrada (Empréstimo e Rendimento), o que equivale a determinar rectas de separação dos dados. Na Figura 2.17 pode ser vista a representação de uma Regressão Linear onde o valor do Empréstimo é considerado como uma função linear do Rendimento.

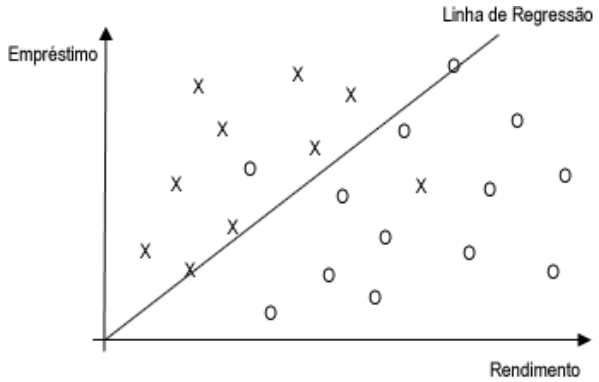


Figura 2.17 - Exemplo de Regressão Linear

O *Clustering* é uma tarefa que permite identificar um conjunto de segmentos ou categorias para descrever os dados. Identifica grupos homogéneos de objectos em que cada grupo é uma classe. O *Clustering* assegura assim que os objectos pertencentes à mesma classe têm um alto grau de similaridade enquanto que os que não pertencem ao mesmo segmento

têm um baixo grau de similaridade. Ao contrário do que acontece com a Classificação, o *Clustering* é uma tarefa de aprendizagem não supervisionada (Han & Kamber, 2001), isto porque o utilizador não tem influência na definição das classes que surgem. Os segmentos são então definidos a partir de agrupamentos que são detectados nos dados e que obedecem a métricas de similaridade.

No *Clustering*, podem ser usados diversos algoritmos para identificar segmentos de dados. As estratégias seguidas podem passar ou pela divisão sucessiva dos registos a segmentar (todos os registos fazem inicialmente parte de um único segmento) ou pela agregação de registos em grupos (inicialmente cada registo representa um segmento) (Berry & Linoff, 2000).

A Figura 2.18 mostra um exemplo de *Clustering* de empréstimos bancários em três segmentos diferentes, em que alguns dos registos pertencem a mais do que um grupo. Segundo Santos e Azevedo (2005), o *Clustering* caracteriza-se pela estimativa da probabilidade de densidade, que consiste em técnicas que estimam a função de probabilidade de densidade de todas as variáveis ou atributos na base de dados.

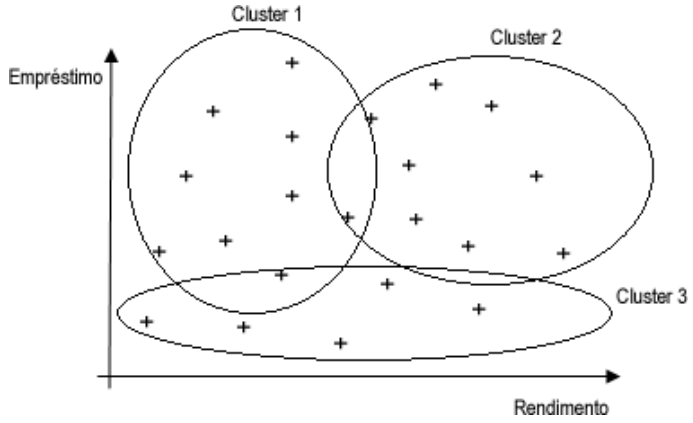


Figura 2.18 - Exemplo de *Clustering*

A tarefa de Associação, ou Dependência, pretende encontrar um modelo que descreva dependências significativas entre variáveis. Esta tarefa permite então identificar grupos de dados tipicamente associados e identificar factos que possam ser directamente ou indirectamente associados, verificando assim a correlação que existe entre os mesmos. As associações que surgem dos dados apresentam um determinado nível de suporte e de confiança que possibilitam a avaliação das relações encontradas.

Fayyad et al. (1996) afirmam que as associações podem surgir a dois níveis:

- Nível estrutural: o modelo é representado de uma forma gráfica e com variáveis localmente dependentes em relação a outras;
- Nível quantitativo: o modelo especifica a força das dependências segundo uma escala numérica.

A Sumariação utiliza métodos para encontrar uma descrição compacta para um subconjunto de dados. Os métodos de Sumariação mais sofisticados derivam de regras de resumo, técnicas de visualizações variadas e as descobertas de relações funcionais entre variáveis. As técnicas de Sumariação são frequentemente aplicadas à análise exploratória de dados e à geração automática de relatórios (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

A Detecção de Desvios, também denominada de Análise de Sequência ou Descoberta de Sequências (Dunham, 2003), tem como foco a descoberta de alterações significativas nos dados. Estas alterações são descobertas a partir de valores medidos previamente ou a partir de valores normativos. Esta tarefa permite descobrir relações nos dados baseadas no tempo. Por exemplo, um determinado conjunto de pessoas vai ao restaurante depois de fazer compras, mas a semana passada foi ao cinema. Segundo Thuraisingham (1999), esta tarefa pode ser aplicada na detecção de fraudes e na detecção de doenças.

Depois de expor as principais tarefas ou objectivos de DM, de seguida vão ser evidenciados os conceitos e as propriedades das técnicas que são mais utilizadas: as AD, a IR e as RNA. Devido ao elevado número de técnicas ou algoritmos de DM existentes (e.g. Algoritmos Genéticos, Redes de Bayes, Máquinas de Vectores de Suporte, *Fuzzy Sets*, *Rough Sets*, *K-Nearest Neighbours*), e por não ser do âmbito deste projecto de Dissertação descrever exaustivamente as diferentes técnicas de DM existentes, apenas são apresentadas as técnicas mais populares evidenciadas anteriormente.

As AD são uma forma de representação de um conjunto de decisões que segue uma hierarquia de classes ou de valores. Como têm uma forma de representação simples, estas são facilmente interpretáveis. As AD são então uma forma de representar conjuntos de regras de classificação.

Os algoritmos de indução de AD constroem as árvores a partir dos dados de treino, subdividindo esses dados até que sejam formados apenas nós puros, ou folhas. As árvores têm a seguinte estrutura (Michalsky, Bratko, & Miroslav, 1998):

- Folhas: ou nós puros, correspondem às classes (objectos);
- Ramos: correspondem aos valores dos atributos;
- Nós internos: correspondem aos atributos a classificar.

A Figura 2.19 mostra um exemplo simples de uma AD onde são inferidas várias regras. O primeiro nó da árvore de decisão é aquele que mostra ser mais correlacionado com a escolha do objecto de saída, que no caso deste exemplo é a concessão ou não de um empréstimo a um determinado indivíduo. Os restantes nós são classificados como estando relacionados com os nós anteriores.

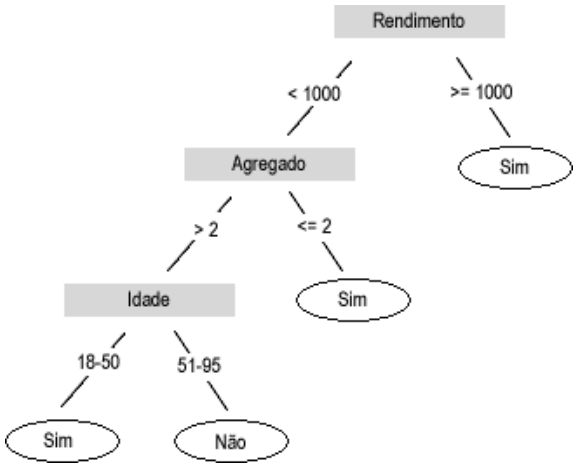


Figura 2.19 - Exemplo de Árvore de Decisão

Existem vários algoritmos para a criação de AD. Os mais conhecidos são o ID3, o C4.5 e o C5, o CART e o CHAID. As suas principais diferenças residem principalmente na forma como determinam os atributos de divisão (e os seus valores), a ordem de aparecimento desses atributos, o número de ramos de cada nó, o critério de paragem e a poda da árvore (*pre-pruning* ou *post-pruning*) (Turban, et al., 2010).

Existem dois tipos de AD, apesar de terem a mesma estrutura. (Berry & Linoff, 2000):

- Árvores de Classificação: qualificam os registos e associa-os com a classe predefinida, garantindo que essa classificação seja correcta;
- Árvores de Regressão: realizam a estimativa do valor de uma determinada variável, utilizando para isso funções matemáticas tais como a média, a mediana ou a moda.

Depois das construções das árvores, estas apresentam uma estrutura ajustada aos dados. Todavia, podem existir ramos desnecessários. A poda da árvore permite produzir árvores menores com melhor potencial e precisão quando novos casos são considerados (Quinlan, 1998), e é por isso um processo importante das AD porque tem como intuito remover partes da árvore que não contribuem para a precisão da classificação, produzindo assim estruturas menos complexas e de mais fácil compreensão. Este procedimento pode ser efectuado de duas maneiras: (1) Durante a aprendizagem (*pre-pruning*), isto é, a árvore é cortada à medida que vai sendo construída. (2) Após a aprendizagem (*post-pruning*), ou seja, a poda é feita depois da identificação da árvore.

Além de poderem ser representadas por árvores, as AD também podem ser apresentadas por conjuntos de regras do tipo "*If…Then…*" que são originadas por cada folha da árvore. Pode ser observada na Figura 2.20 uma regra associada à AD apresentada na Figura 2.19.

```
IF (Rendimento < 1000) AND (Agregado > 2) AND (Idade = "51-95")
THEN "Não"
```

Figura 2.20 - Exemplo de uma regra de uma Árvore de Decisão

A IR, ou Regras de Associação, é uma técnica de DM que permite detectar tendências e padrões em grupos de dados, ou seja, regras sobre os dados. Geralmente, essas regras são apresentadas em forma de lista (Berson, Smith, & Thearling, 2000).

Segundo Quinlan (1998), o objectivo desta técnica é encontrar padrões nos dados, isto é, dependências entre os atributos ou os valores, através da análise das probabilidades condicionais. Normalmente, os resultados aparecem sob a forma de regras $X \rightarrow Y$ (se X está presente, então Y também tem probabilidade de estar presente), em que X pode ser um conjunto de atributos e valores e Y representa um único atributo não presente em X. As

regras têm dois graus associados, a confiança e o suporte. A confiança é a própria probabilidade condicional da regra, isto é, é obtida pela divisão do número de casos onde X e Y aparecem juntos pelo número de casos onde apenas aparece o X. Por sua vez, o suporte é o número de ocorrências desta regra nos dados, ou seja, o número de vezes que X está associado a Y. É importante analisar estas duas medidas simultaneamente, isto porque o suporte (registos que satisfazem a regra) de uma determinada regra pode ser elevado, mas a confiança (número de registos em que é possível prever X conhecendo Y) da mesma ser diminuta.

A Figura 2.21 apresenta um exemplo de uma associação. A regra $Televisão\ e\ Leitor\ de\ DVD \rightarrow Portátil$ indica que os clientes que compraram o produto $Televisão$ juntamente com o produto $Leitor\ de\ DVD$ também compraram o produto $Portátil$. A regra tem uma confiança de 0.66 (66%), visto que foi verificada a ocorrência da compra dos produtos $Televisão$ e $Leitor\ de\ DVD$ em 3 casos, mas apenas em 2 é que também foi confirmada a compra do produto $Portátil$. O suporte da regra é de 50%, uma vez que se verifica a regra em metade dos clientes.

Cliente	Produto
1	Portátil
1	Impressora
1	Televisão
1	Leitor de DVD
2	Televisão
2	Leitor de DVD
2	DVD
3	Portátil
3	Tablet
4	Televisão
4	Leitor de DVD
4	Portátil

X: Televisão e Leitor de DVD
Y: Portátil

Confiança: (2/3) 0.66
Suporte: 2 (50%)

Figura 2.21 - Exemplo de Indução de Regras

As RNA, assim como outras técnicas de DM, como os algoritmos genéticos, têm a sua base na biologia. Esta técnica, mais precisamente, tem um funcionamento análogo ao processamento do cérebro humano, sendo constituída por conjuntos de neurónios (ou nodos). Groth (2000) afirma que cada neurónio recebe uma série de valores e, em função destes, apresenta um valor de saída. Estes valores de saída, nalguns neurónios, são também os valores de entrada de outros, dependendo das interligações existentes entre os nodos. A maneira como as interligações são estabelecidas é importante para o tipo de resultados a obter.

As principais características das RNA são (Hagan, Demuth, & Beale, 1996):

- Flexibilidade: Têm um grande domínio de aplicabilidade;
- Adaptabilidade: Podem adaptar a sua topologia de acordo com mudanças do ambiente;
- Processamento paralelo: Permitem que tarefas complexas sejam realizadas num curto espaço de tempo;
- Aprendizagem e generalização: Conseguem descrever o todo a partir de algumas partes, estabelecendo-se como formas eficientes de armazenamento de conhecimento e de aprendizagem;
- Robustez e degradação suave: Permitem processar a informação incompleta, ou ruído, de uma maneira eficiente e são capazes de manter o seu desempenho quando acontece a desactivação de alguns neurónios e/ou conexões.

Para se construir uma RNA é preciso determinar o número de neurónios, definir o seu tipo, como é que estes vão estar ligados, definir o processo de codificação da entrada e descodificação da saída, iniciar os pesos da rede e proceder ao treino da rede por aplicação de um algoritmo (Groth, 2000).

O neurónio é o elemento chave para as operações das RNA. Este é constituído por três elementos fundamentais (Hagan, et al., 1996); (Cortez, 2002):

- Um conjunto de ligações que representam as sinapses, ou ligações entre neurónios. Cada conexão tem um peso associado (w_{ij}) que representa a força do sinal enviado, que tem um efeito de excitação (peso positivo) e um efeito de inibição (peso negativo). Assim, o sinal de entrada (x_j) é multiplicado pelo peso correspondente (w_{ij}), onde i representa o neurónio objecto de estudo e j o neurónio emissor do sinal. Existe também geralmente uma conexão extra que adiciona um valor constante (1) para que se estabeleçam as correctas condições operacionais para o neurónio;
- Um integrador, círculo que representa o neurónio, que reduz os n argumentos de entrada (estímulos) a um único valor de saída. É utilizada a função adição ($\sum$), pesando todas as entradas numa combinação linear;
- Uma função de activação (fa) que limita a amplitude da saída do neurónio, introduzindo uma componente não linear.

A Figura 2.22, adaptada de Quintela (2005) representa um neurónio artificial, sendo que a soma pesada das entradas do neurónio é descrito por $U_i = \sum(w_{i0}, x_1 \times w_{i1}, x_2 \times w_{i2}, \dots, x_n \times w_{in})$ e o resultado do cálculo do valor de activação do neurónio usando a função de activação é descrito por $S_i = fa(U_i)$.

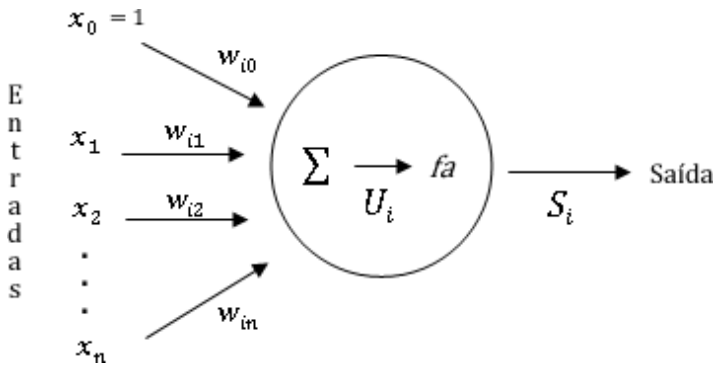


Figura 2.22 - Neurónio Artificial

Existem três funções de activação mais recorrentes (Hagan, et al., 1996; Kartalopoulos, 1996):

- *Step*: função linear representada pela função $f(x) = \begin{cases} 1, & x \geq t \\ 0, & x < t \end{cases}$
- *Sign*: função linear representada pela função $f(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$
- Sigmóide: função não linear representada pela função $f(x) = \frac{1}{1+e^{-x}}$

2.2 A Doença Pulmonar Obstrutiva Crónica

2.2.1 Definição

A DPOC caracteriza-se por ser uma doença respiratória crónica progressiva que é possível prevenir e tratar. A limitação das vias aéreas nunca é plenamente corrigível, geralmente progressiva e associada a uma resposta inflamatória anormal a partículas e gases agressivos (GOLD, 2010).

Quando é diagnosticada a DPOC, há geralmente duas doenças associadas: Bronquite Crónica e Enfisema (SPP, 1997). A Bronquite crónica caracteriza-se pelo estreitamento das vias aéreas por causa da hipertrofia dos músculos da árvore brônquica e pela produção excessiva de muco, suficiente para causar expectoração excessiva (na maioria dos dias durante pelo

menos 3 meses no ano durante pelo menos 2 anos sucessivos). O Enfisema caracteriza-se pelo aumento dos espaços aéreos dos bronquíolos por destruição das suas paredes, não substituída por fibrose.

Não são englobadas na DPOC outras patologias que se podem acompanhar de obstrução das vias aéreas, nomeadamente a asma (que apresenta reversibilidade da obstrução com períodos de normalidade e tem maior variabilidade dos débitos expiratórios), as bronquiectasias, a fibrose quística, a bronquiolite e a obstrução das vias aéreas superiores. De igual modo, nem todos os doentes com Bronquite Crónica ou Enfisema têm DPOC (SPP, 1997).

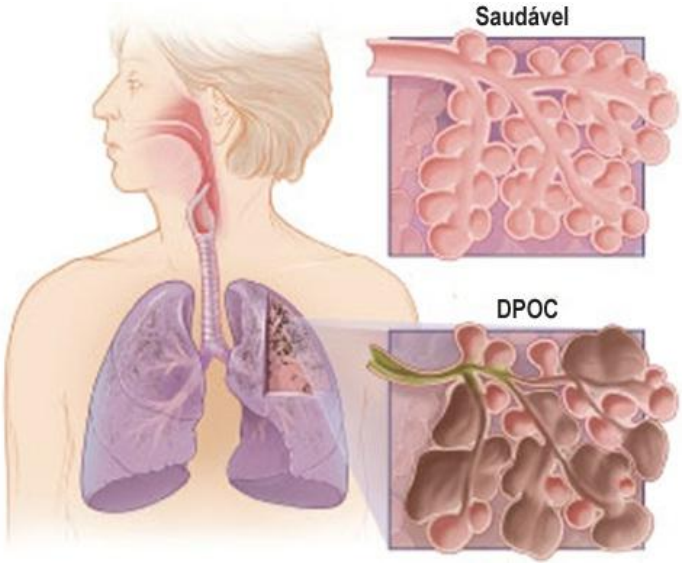


Figura 2.23 - Pulmão saudável vs pulmão com DPOC

A Figura 2.23, adaptada de Vivacare (2010), mostra a diferença entre um pulmão saudável e um pulmão com DPOC.

2.2.2 Factores de Risco

O consumo de tabaco tem provado ser o maior factor de risco ligado a esta doença. Todavia, outros factores têm vindo a ser associados à DPOC, a maioria sem serem comprovados cientificamente (GOLD, 2010).

- Exposições inalatórias: Fumo do tabaco (fumador activo ou passivo), poeiras e exposições químicas relacionadas com a profissão e poluição (interior e exterior);

- Sexo, Idade e Raça: A doença parece ser mais prevalente no sexo masculino e em idades avançadas e a mortalidade por DPOC parece ser mais elevada na raça caucasiana;
- Estatuto Socioeconómico: Indivíduos provenientes de estratos sociais mais baixos têm maior risco de desenvolver DPOC, embora os factos que promovem esta associação ainda não sejam totalmente claros;
- Infecções respiratórias: infecções como a pneumonia ou a tuberculose poderão estar relacionadas com a DPOC;
- Condições genéticas: A única alteração genética conhecida que condiciona o aparecimento de DPOC é a deficiência severa de 1-antitripsina que se relaciona com o aparecimento prematuro de Enfisema e geralmente acompanhado de bronquite crónica ou bronquiectasias.

2.2.3 Diagnóstico

Existem vários aspectos a ter em consideração para o diagnóstico de DPOC. Os mais importantes são os seguintes (Vivacare, 2010):

- Sintomas: Os sintomas mais importantes associados a esta doença são a presença de dispneia, tosse crónica e/ou a existência permanente (e excesso) de produção de muco;
- Historial médico: Anteriores exposições a factores de risco da doença e/ou hospitalizações relacionadas com problemas respiratórios;
- Exame físico: Também pode ajudar na identificação da doença, mas os sinais só aparecem num estado avançado. Um exemplo de um sinal pode ser o "tórax em barril" (Loveridge, West, Kryger, & Anthonisen, 1986);
- Espirometria - Medição da limitação do fluxo das vias aéreas: Este teste é dividido em duas etapas. A primeira fase é a medição do FVC (*Forced Vital Capacity*) que é a medição da expiração de ar total a partir do ponto inicial de inspiração máxima. A segunda fase é a medição do ar expirado na primeira etapa, mas apenas durante o primeiro segundo da expiração, denominada de FEV1 (*Forced Expiration Volume In One Second*). Depois de se obter este dois valores, o rácio FEV1/FVC tem que ser calculado. A Figura 2.24, adaptada de McCarthy & Dweik (2010), ilustra o padrão normal de uma espirometria e um possível padrão anormal.

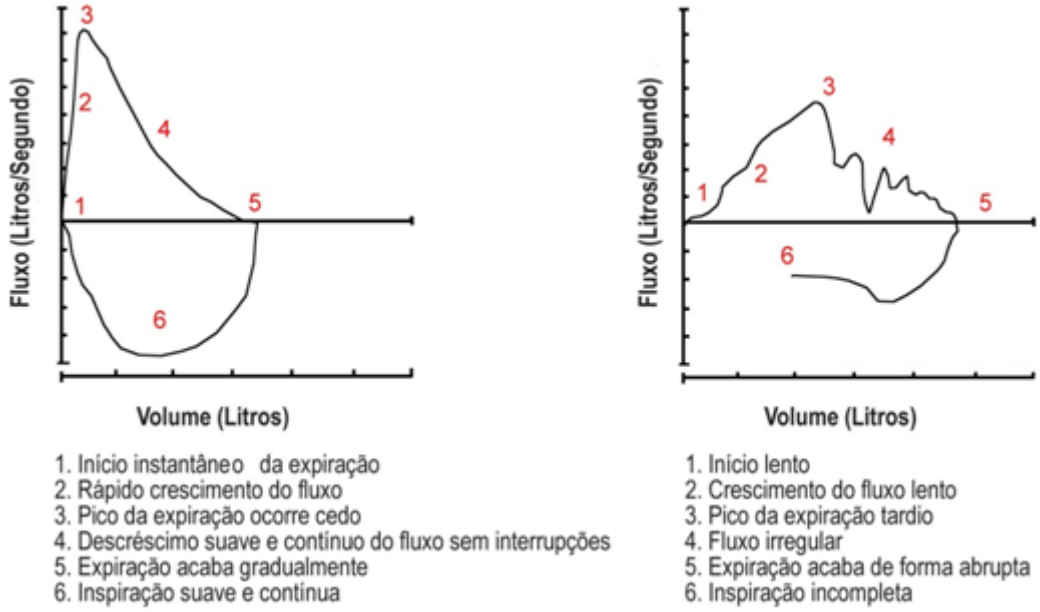


Figura 2.24 - Características do fluxo respiratório normal e anormal numa espirometria

2.2.4 Níveis de gravidade

Esta doença está dividida em quatro níveis de gravidade. A Tabela 2.4 mostra os valores padrão dos quatro níveis. É importante referir que para medir o nível de gravidade da DPOC, os valores obtidos na espirometria devem ser comparados de acordo com a idade, altura, sexo e raça. É de realçar que esta classificação é baseada em espirometrias feitas com o uso prévio de broncodilatadores (GOLD, 2010).

Nível 1 - Ligeiro	FEV1/FVC < 0,70 FEV1 ≥ 80% previsto
Nível 2 - Moderado	FEV1/FVC < 0,70 50% ≤ FEV1 < 80% previsto
Nível 3 - Grave	FEV1/FVC < 0,70 30% ≤ FEV1 < 50% previsto
Nível 4 – Muito Grave	FEV1/FVC < 0,70 FEV1 < 30% previsto ou FEV1 < 50% previsto mais falhas respiratórias crónicas

Tabela 2.4 - Níveis de gravidade da DPOC

Capítulo 3 – O Sistema de *Business Intelligence* para o estudo da Doença Pulmonar Obstrutiva Crónica

Neste capítulo é retratado o sistema de BI concebido e implementado. Este capítulo começa pela apresentação das características dos dados disponíveis. Também é descrita neste capítulo a arquitectura do sistema assim como as várias partes da sua implementação, desde a BDO da aplicação *Web*, assim como uma apresentação das principais funcionalidades da própria aplicação, até à concepção e implementação do DW que servirá de suporte às análises apresentadas no capítulo seguinte.

3.1 Caracterização dos dados disponíveis

Este projecto de dissertação, como referido anteriormente, resulta do trabalho de investigação e desenvolvimento que tem vindo a ser realizado entre o Departamento de Sistemas de Informação da Universidade do Minho e a organização sem fins lucrativos FPP. Esta última realiza várias actividades/iniciativas para recolher, guardar e analisar dados referentes a diversas doenças respiratórias. Os resultados obtidos são utilizados para caracterizar a realidade actual do país e para planear e preparar acções que têm como intuito melhorar a qualidade de vida dos cidadãos.

O conjunto de dados disponível para este projecto de dissertação foi recolhido pela FPP em iniciativas realizadas em Portugal durante o ano de 2007. Estas iniciativas são públicas e portanto abertas a todos aqueles que desejam participar. Nestas, os participantes são convidados a responder a um questionário que integra questões relacionadas com os sintomas e factores de risco da DPOC. O questionário inclui também informações sobre a localização geográfica da residência do paciente, assim como o seu sexo, idade, peso e altura. Para além dos dados referidos acima, os dados do exame espirométrico são igualmente registados. O conjunto de dados disponibilizado apresenta 32 atributos e 1898 registos, estando numa folha

do *Microsoft Excel*. De seguida, são apresentados na Tabela 3.1 os atributos presentes nessa folha de cálculo, a sua descrição e os valores que cada campo deve apresentar.

Designação Atributo	Descrição e valores possíveis
Idade	Idade do utente (número inteiro)
Sexo	Sexo do utente {m, f}
Localidade	Descrição da localidade (nome de localidade)
Peso	Peso do utente (número inteiro)
Altura	Altura do utente (número inteiro)
IMC	Índice de Massa Corporal do utente (número decimal)
Tabaco	Resposta às perguntas do inquérito "É fumador? É ex-Fumador?" {"s": sim, "n": não, "ex": ex-fumador}
Nariz / Espirros	Resposta à pergunta do inquérito "Já algumas vezes teve crises de espirros, corrimento nasal ou nariz entupido, quando não está constipado ou com gripe?" {s, n}
Lacrimejo / Comichão	Resposta à pergunta do inquérito "Nos últimos 12 meses esse problema do nariz foi acompanhado de olhos lacrimejantes e com comichão?" {s, n}
Tosse seca	Resposta à pergunta do inquérito "Tem períodos de tosse seca, irritativa (alérgica)?" {s, n}
Tosse +3 meses	Resposta à pergunta do inquérito "Tosse muitas vezes na maior parte dos dias? (se somados todos os dias acha que ultrapassa os três meses?)" {s, n}
Expectoração diária	Resposta à pergunta do inquérito "Tem expectoração ou "catarro" na maioria dos dias?" {s, n}
Pieira	Resposta à pergunta do inquérito "Já alguma vez teve pieira ou assobios (chieira, apertos no peito)?" {s, n}
Pieira últimos 12 meses	Resposta à pergunta do inquérito "Isso aconteceu nos últimos 12 meses?(pieira)" {s, n}
Pieira com gripe	Resposta à pergunta do inquérito "Isso aconteceu quando estava constipado com gripe? (pieira)" {s, n}
Cansaço	Resposta à pergunta do inquérito "Cansa-se mais do que as outras pessoas da sua idade?" {s, n}
Falta de ar	Resposta à pergunta do inquérito "Tem que parar com "falta de ar" ao subir uma ladeira ou um lanço de escada?" {s, n}
Alergias	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de alergias?" {s, n}
Rinite	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de rinite alérgica?" {s, n}
Asma	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de asma brônquica?" {s, n}
Medicação asma diária	Resposta à pergunta do inquérito "Faz medicação para a asma? (diariamente)" {s, n}
Medicação asma crises	Resposta à pergunta do inquérito "Faz medicação para a asma? (só em crises)" {s, n}
DPOC	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de bronquite crónica, DPOC (Doença Pulmonar Obstrutiva Crónica) ou enfisema?" {s, n}
Gripe +2xano	Resposta à pergunta do inquérito "Constipa-se ou tem gripe mais do que duas vezes no ano?" {s, n}
Pneumonia	Resposta à pergunta do inquérito "Já teve pneumonia ou broncopneumonia?" {s, n, ns}
TB	Resposta à pergunta do inquérito "Já teve tuberculose pulmonar? Quando?" {s, n, ns} (Idade[0-120])
BCG	Resposta à pergunta do inquérito "Está vacinado com a BCQ (vacina contra a Tuberculose)?" {s, n, ns}
Vacina Gripe	Resposta à pergunta do inquérito "Vacina-se habitualmente contra a gripe?" {s, n, ns}
Vacina Pneumonia	Resposta à pergunta do inquérito "Está vacinado contra a pneumonia?" {s, n, ns}
Outras Vacinas	Resposta à pergunta do inquérito "Faz habitualmente outras "vacinas" para prevenção das infecções respiratórias?" {s, n, ns}
FEV1	Valor percentual do FEV1
FEF 25-75	Valor percentual do FEF 25-75

Tabela 3.1 – Tabela resumo dos dados disponíveis

3.2 Arquitectura do Sistema

Nesta subsecção são apresentadas tanto a arquitectura do sistema proposto, como as tecnologias que serão utilizadas para a sua implementação.

Os dados fornecidos pela FPP, como referido anteriormente, são disponibilizados numa folha de cálculo do *Microsotf Office Excel* e portanto, será a partir destes dados que será efectuado todo o processo de ETL que permitirá a migração dos mesmos da folha de cálculo para o DW. Este DW servirá de suporte à tecnologia OLAP, que analisa os dados sobre as diferentes perspectivas consideradas na modelação do DW, e aos algoritmos de DM, que integram conceitos como a estatística e a inteligência artificial para identificar modelos, padrões ou tendências nos dados. Tanto o DW como o OLAP e o DM serão implementados / realizados utilizando a tecnologia do *Microsotf SQL Server 2008*.

Para os inquéritos e dados futuros da FPP, foi implementada uma aplicação *Web* (utilizando tecnologias como as páginas cliente/servidor .ASP, e as linguagens de programação HTML e Javascript) para facilitar o processo de recolha e gestão dos dados iniciais de todo este sistema de BI, uma vez que os seus dados serão armazenados numa BDO. Esta BDO, também implementada com a tecnologia do *Microsotf SQL Server 2008*, tornará o processo de ETL e a respectiva migração dos dados para o DW mais eficaz e mais eficiente. Portanto, as folhas de cálculo e os inconvenientes que estas acarretam para este tipo de sistema serão substituídas pela aplicação *Web* desenvolvida que melhorará não só todo o processo de ETL mas também a qualidade dos dados recolhidos.

A arquitectura do sistema de BI proposto descrita precedentemente é ilustrada na Figura 3.1.

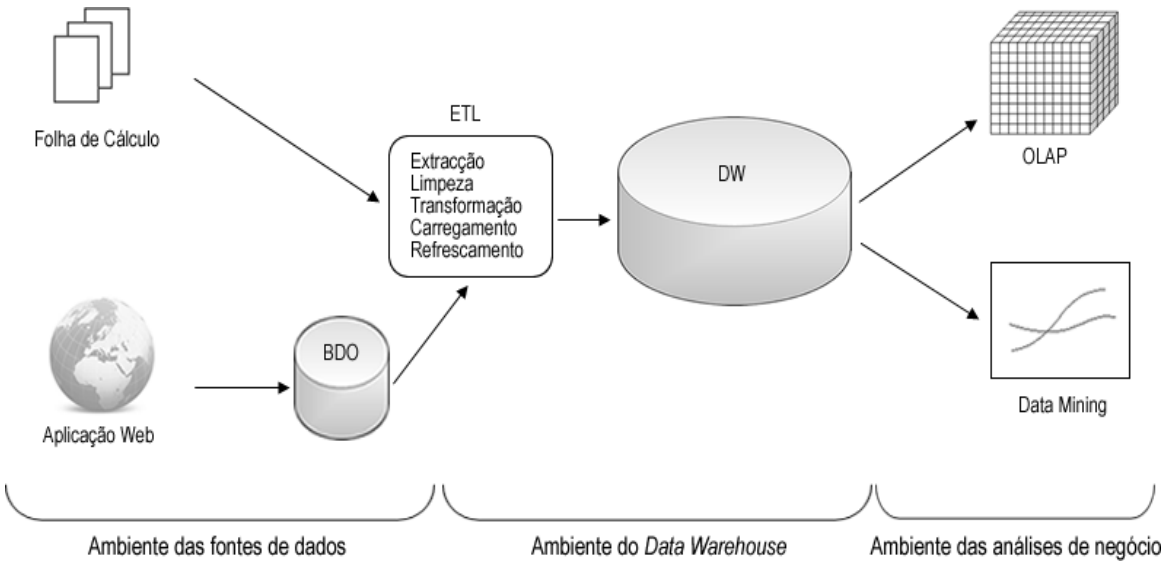


Figura 3.1 - Arquitectura do sistema implementado

3.3 Implementação do Sistema

3.3.1 Aplicação *Web*

Esta subsecção tem como objectivo evidenciar o que foi efectivamente desenvolvido do que inicialmente estava previsto ser apenas um protótipo, mas que agora é uma aplicação *Web* completamente funcional e utilizada pela FPP.

Começando pela BDO que suporta os dados da aplicação, esta contém 12 tabelas. De seguida é feita uma breve descrição de cada uma dessas tabelas que compõem a BDO da aplicação, apresentando os atributos que a compõem assim como o seu tipo de dados e respectiva descrição.

- **Utente**: Contém todos os dados pessoais referentes ao Utente, como por exemplo o seu nome, idade, peso, altura mas também dados relativos à localização da sua residência e profissão (Tabela 3.2);

Utente				
Campo	**Tipo**	**Chave**		**Descrição**
		P[3]	**E[4]**	
Id_Utente	int	X		Número de Identificação do Utente
NomeEncr	varbinary (max)			Nome Encriptado do Utente
Dia_Nasc	int			Dia de nascimento do Utente
Mes_Nasc	Int			Mês de nascimento do Utente
Ano_Nasc	Int			Ano de nascimento do Utente
Sexo	nvarchar (50)			Sexo (M ou F) do Utente
Peso	Int			Peso (em kg) do Utente
Altura	Int			Altura (em cm) do Utente
Morada	nvarchar (max)			Morada do Utente
CP4	Int		X	Primeiros quatro dígitos do Código Postal do Utente
CP3	Int		X	Últimos três dígitos do Código Postal do Utente
Id_ProfissaoN3	nvarchar (50)		X	Número de Identificação da Profissão do Utente

Tabela 3.2 - Tabela "Utente" da BDO da aplicação *Web*

- **CodigoPostal**: Contém todos os códigos postais de Portugal e respectivas freguesias associadas (Tabela 3.3);

CodigoPostal				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
CP4	Int	X		Primeiros quatro dígitos do Código Postal
CP3	Int	X		Últimos três dígitos do Código Postal
Id_Freguesia	nvarchar (50)		X	Número de Identificação da Freguesia

Tabela 3.3 - Tabela "CodigoPostal" da BDO da aplicação *Web*

- **Freguesia:** Contém todas as freguesias de Portugal e respectivos concelhos associados (Tabela 3.4);

Freguesia				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Freguesia	nvarchar (50)	X		Número de Identificação da Freguesia
Id_Concelho	nvarchar (50)		X	Número de Identificação do Concelho
Designacao	nvarchar (50)			Designação da Freguesia

Tabela 3.4 - Tabela "Freguesia" da BDO da aplicação *Web*

[3] P: Chave Primária (*Primary Key*).

[4] E: Chave Estrangeira (*Foreign Key*).

- **Concelho:** Contém todos os concelhos de Portugal e respectivos distritos associados (Tabela 3.5);

Concelho				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Concelho	nvarchar (50)	X		Número de Identificação do Concelho
Id_Distrito	nvarchar (50)		X	Número de Identificação do Distrito
Designacao	nvarchar (50)			Designação do Concelho

Tabela 3.5 - Tabela "Concelho" da BDO da aplicação *Web*

- **Distrito**: Contém todos os distritos de Portugal (Tabela 3.6);

Distrito				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Distrito	nvarchar (50)	X		Número de Identificação do Distrito
Designacao	nvarchar (50)			Designação do Distrito
Cod_Distrito	int			Código do Distrito

Tabela 3.6 - Tabela "Distrito" da BDO da aplicação *Web*

No que diz respeito à identificação da profissão do Utente, as próximas três tabelas seguem a mesma lógica da primeira hierarquia mostrada acima (Código Postal -> Freguesia -> Concelho -> Distrito), sendo as profissões agrupadas segundo o modelo da Classificação Nacional de Profissões[5]. Todavia, não sendo necessário esmiuçar as profissões até ao nível mais baixo da hierarquia (até porque complicaria a inserção destes dados no inquérito, demorando imenso tempo apenas para esta questão), achou-se suficiente implementar os três primeiros níveis desta classificação. É de notar ainda que, para além das profissões presentes na Classificação Nacional de Profissões, e porque seria necessário devido ao âmbito desta dissertação e das iniciativas da FPP, acrescentou-se os seguintes estatutos às "profissões": Aposentados/Reformados; Estudantes; Desempregados; Grupos de Risco (Prostitutas/os, Encarcerados, Toxicodependentes, Sem-Abrigo).

[5] http://www.iefp.pt/formacao/CNP/Paginas/CNP.aspx

- **ProfissaoN3**: Contém todas as profissões do 3º nível (nível mais baixo) da hierarquia (Tabela 3.7);

ProfissaoN3				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_ProfissaoN3	nvarchar (50)	X		Número de Identificação da Profissão de Nível 3
Id_ProfissaoN2	nvarchar (50)		X	Número de Identificação da Profissão de Nível 2
Designacao	nvarchar (max)			Designação da Profissão de Nível 3

Tabela 3.7 - Tabela "ProfissaoN3" da BDO da aplicação *Web*

- **ProfissaoN2**: Contém todas as profissões do 2º nível da hierarquia (Tabela 3.8);

ProfissaoN2				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_ProfissaoN2	nvarchar (50)	X		Número de Identificação da Profissão de Nível 2
Id_ProfissaoN1	nvarchar (50)		X	Número de Identificação da Profissão de Nível 1
Designacao	nvarchar (max)			Designação da Profissão de Nível 2

Tabela 3.8 - Tabela "ProfissaoN2" da BDO da aplicação *Web*

- **ProfissaoN1**: Contém todas as profissões do 1º nível da hierarquia (Tabela 3.9);

ProfissaoN1				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_ProfissaoN1	nvarchar (50)	X		Número de Identificação da Profissão de Nível 1
Designacao	nvarchar (max)			Designação da Profissão de Nível 1

Tabela 3.9 - Tabela "ProfissaoN2" da BDO da aplicação *Web*

- **QuestoesRespostas**: Esta tabela contém as respostas dos utentes ao inquérito da aplicação *Web* e é portanto outra das tabelas principais desta base de dados, juntamente com a tabela "Utente" (Tabela 3.10);

QuestoesRespostas				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Utente	Int	X		Número de Identificação do Utente
Id_Questao	nvarchar (10)	X		Número de Identificação da Questão
Id_Utilizador	Int		X	Número de Identificação do Utilizador
Resposta	nvarchar (max)			Resposta à Questão
Dia	Int	X		Dia do preenchimento do Inquérito
Mes	Int	X		Mês do preenchimento do Inquérito
Ano	Int	X		Ano do preenchimento do Inquérito
DataC	Date			Data Completa do preenchimento do Inquérito
Idade	Int			Idade do Utente aquando do preenchimento do Inquérito
Id_Populacao	Int		X	Número de Identificação da População do Utente aquando do preenchimento do Inquérito

Tabela 3.10 - Tabela "QuestoesRespostas" da BDO da aplicação *Web*

- **Questao**: – Nesta tabela estão armazenadas as perguntas, e as questões relativas à espirometria, também elas incluídas no inquérito (Tabela 3.11);

Questao				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Questao	nvarchar (10)	X		Número de Identificação da Questão
Designacao	nvarchar (50)			Designação da Questão

Tabela 3.11 - Tabela "Questao" da BDO da aplicação *Web*

- **Populacao**: Esta tabela contém as diferentes "populações" específicas onde se faz o inquérito, como por exemplo "*British Hospital*", "2ª Feira do Pulmão", etc. (Tabela 3.12);

Populacao				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Populacao	Int	X		Número de Identificação da População
Descricao	nvarchar (max)			Descrição da População

Tabela 3.12 - Tabela "Populacao" da BDO da aplicação *Web*

- **Utilizador**: Esta tabela contém os dados dos utilizadores autorizados a aceder à aplicação *Web* (Tabela 3.13).

Utilizador				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Id_Utilizador	Int	X		Número de Identificação do Utilizador
Username	nvarchar (50)			Username do Utilizador
Nome	nvarchar (50)			Nome do Utilizador
Email	nvarchar (50)			E-mail do Utilizador
Tipo	nvarchar (50)			Tipo (admin, med ou vol)[6] de Utilizador
PassEncr	varbinary (max)			Password Encriptada do Utilizador

Tabela 3.13 - Tabela "Utilizador" da BDO da aplicação *Web*

Na Figura 3.2 é apresentado o Diagrama de Entidades e Relacionamentos (DER) da BDO que ilustra as descrições das tabelas.

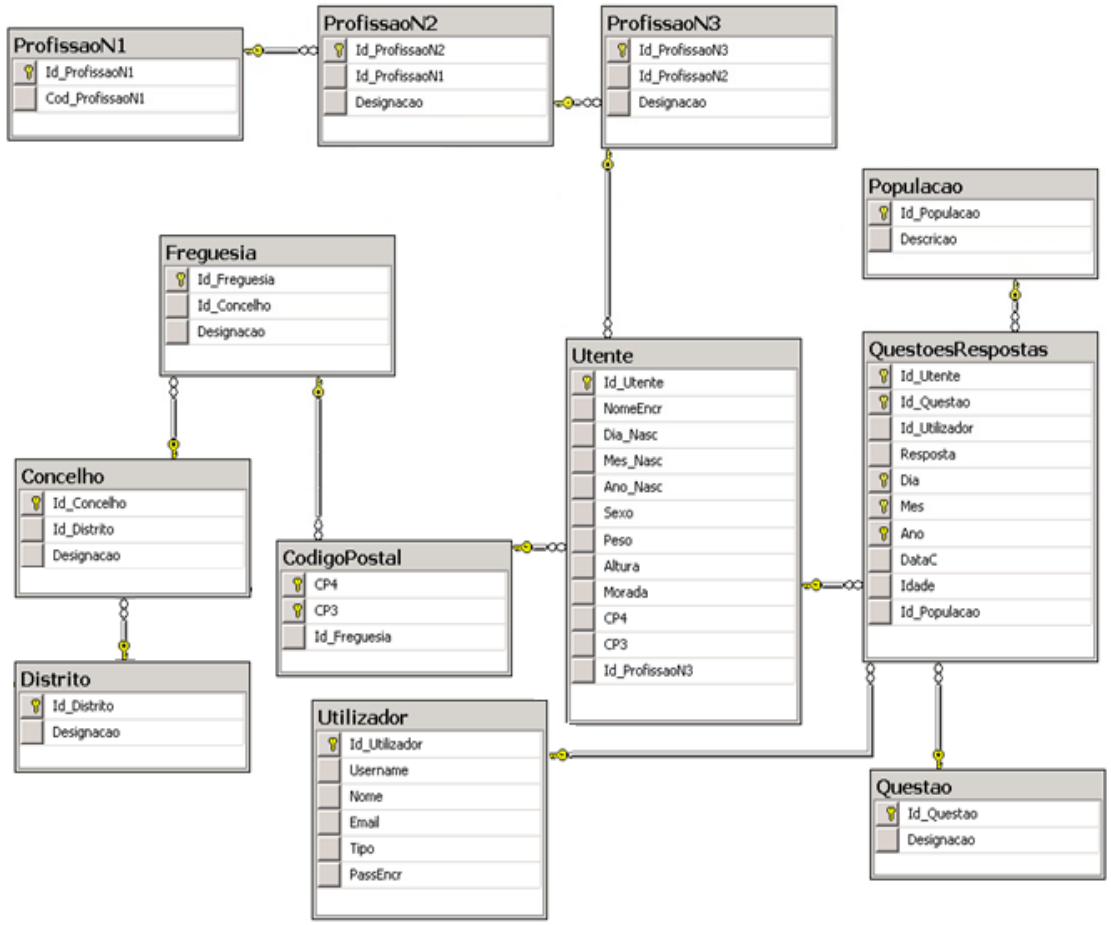


Figura 3.2 - DER da BDO da Aplicação *Web*

[6] Os tipos de Utilizador da aplicação *Web* podem ser os seguintes: “admin” (Administrador), “med” (Médico) e “vol” (Voluntário)

De referir ainda que tanto o nome do utente como a *password* do utilizador se encontram encriptados na base de dados do *Microsoft SQL Server 2008* por questões de ética/confidencialidade e de segurança, respectivamente. Essa encriptação foi feita utilizando o *Service Master Key* do próprio *Microsoft SQL Server 2008* e *symmetric keys* usando o algoritmo AES_256, também nativo do *Microsoft SQL Server 2008.*

Após a breve descrição da base de dados que suporta a aplicação *Web*, de seguida são mostradas e explicadas algumas das suas funcionalidades.

A primeira página que o utilizador visualiza é a página de autenticação, ilustrada pela Figura 3.3. Se tentar visualizar as outras páginas sem se autenticar, não só não consegue ver os seus conteúdos como é automaticamente reencaminhado para esta página de *login*.

Figura 3.3 - *Login*

O Utilizador deve introduzir o "Username" e a "Password" que lhe foram atribuídos (atenção que o sistema diferencia as letras minúsculas de maiúsculas) e de seguida carregar no botão circundado a vermelho.

Depois de efectuada a sua autenticação, o Utilizador depara-se com um menu que contém três (se for um médico: "Novo", "Pesquisar", "Os meus Inquéritos"; se for um administrador: "Novo", "Pesquisar", "Gerir Tabelas".) ou duas escolhas (se for um voluntário: "Novo", "Pesquisar"). Esta situação é representada pela Figura 3.4, após a autenticação de um administrador.

Figura 3.4 - *Home*

Caso seja a primeira vez que o utente participa a estes inquéritos via a aplicação *Web*, o Utilizador deverá carregar no botão "Novo", caso contrário (ou se tiver dúvidas) deve carregar no botão "Pesquisar" para verificar a existência prévia do utente no sistema.

Ao seleccionar "Novo", deverá aparecer uma imagem semelhante à Figura 3.5.

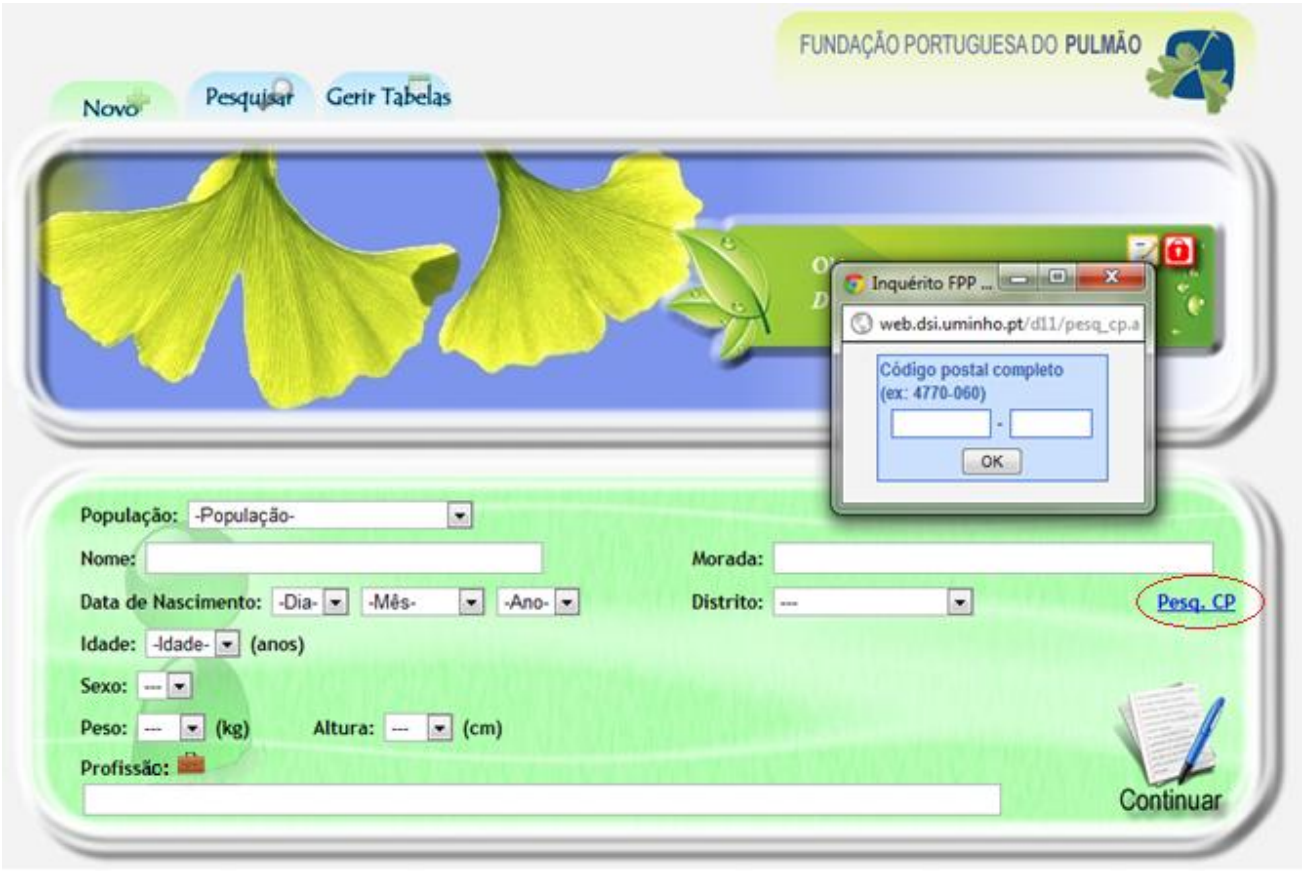


Figura 3.5 - Dados do Utente (1)

Os dados do Utente devem ser devidamente preenchidos. O Utilizador apenas terá de preencher os campos que aparecem e escolher algumas opções que já vêm limitadas e com as respectivas validações (e.g. População, Data de Nascimento, Sexo, Peso, Altura, Distrito, Concelho, Freguesia, Código Postal e Profissão). De notar ainda que o campo Idade é preenchido automaticamente quando a Data de Nascimento é alterada. Este último campo é o único que não é de preenchimento obrigatório. Quando o utilizador escolher o distrito, apenas os respectivos concelhos aparecerão, e assim sucessivamente até à escolha dos códigos postais de uma determinada freguesia (a janela de pesquisa por código postal só aparece quando se clica no "Pesq. CP" e é útil para obter o Distrito/Concelho/Freguesia quando o utente apenas sabe o seu código postal).

Tal como na selecção do Distrito/Concelho/Freguesia/Código Postal, a selecção da profissão é feita com o aparecimento de uma janela para a escolha da mesma através da selecção das várias hierarquias. Aquando da escolha da profissão (), um ficheiro que auxilia esse processo estará sempre disponível, carregando no ícone de ajuda da respectiva janela (Figura 3.6). Deverá ser seleccionada a <u>última profissão</u> realizada pelo utente. Apenas se deve seleccionar a opção "Estudante" se o utente ainda não tiver concluído o seu ciclo de estudos e "Desempregado" se o utente nunca teve nenhum emprego.

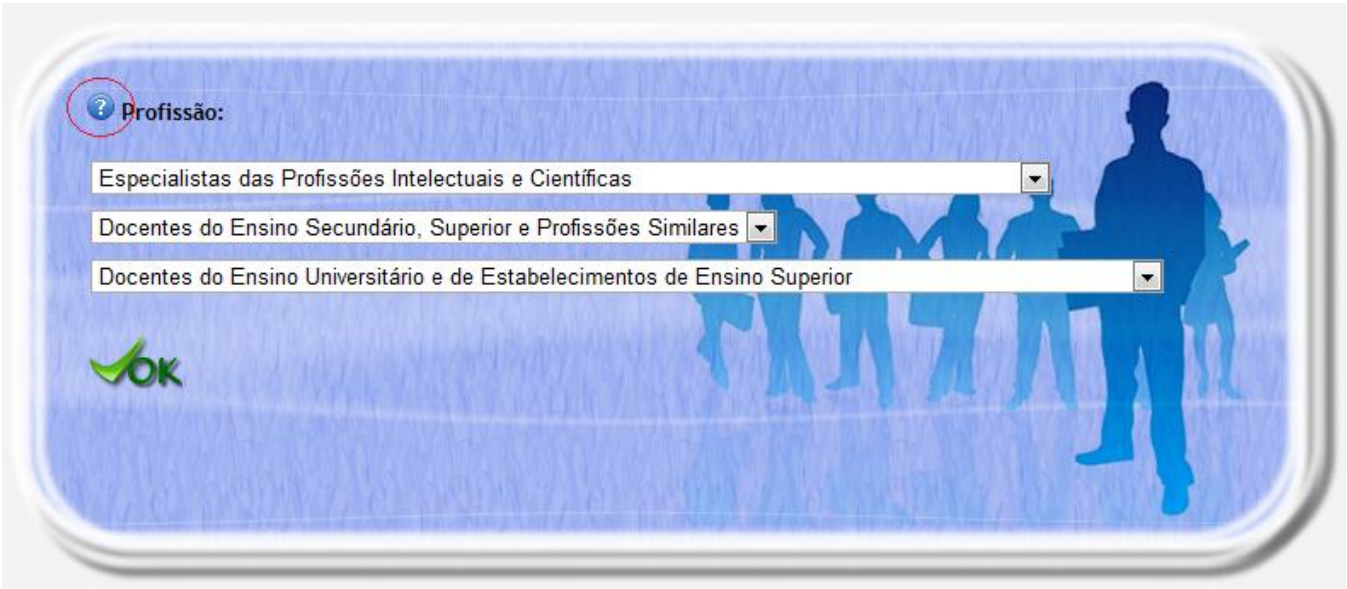


Figura 3.6 - Escolha da Profissão

Finalizado o processo de preenchimento dos dados do Utente, o resultado deverá ser semelhante ao apresentado na Figura 3.7. Carregar em "Continuar" para preencher o resto do Inquérito.

Figura 3.7 - Dados do Utente (2)

O Utilizador deve preencher devidamente o inquérito (Figura 3.8 e Figura 3.9) segundo as respostas do Utente. No final do inquérito tem de preencher a área referente aos dados retirados da espirometria assim como, se necessário (opcional), notas sobre o utente. A selecção da Bronco Motricidade é também opcional, permitindo o botão "Limpar" remover a opção seleccionada. O questionário já está devidamente implementado com as validações necessárias, não deixando por exemplo que alguma questão seja deixada sem resposta.

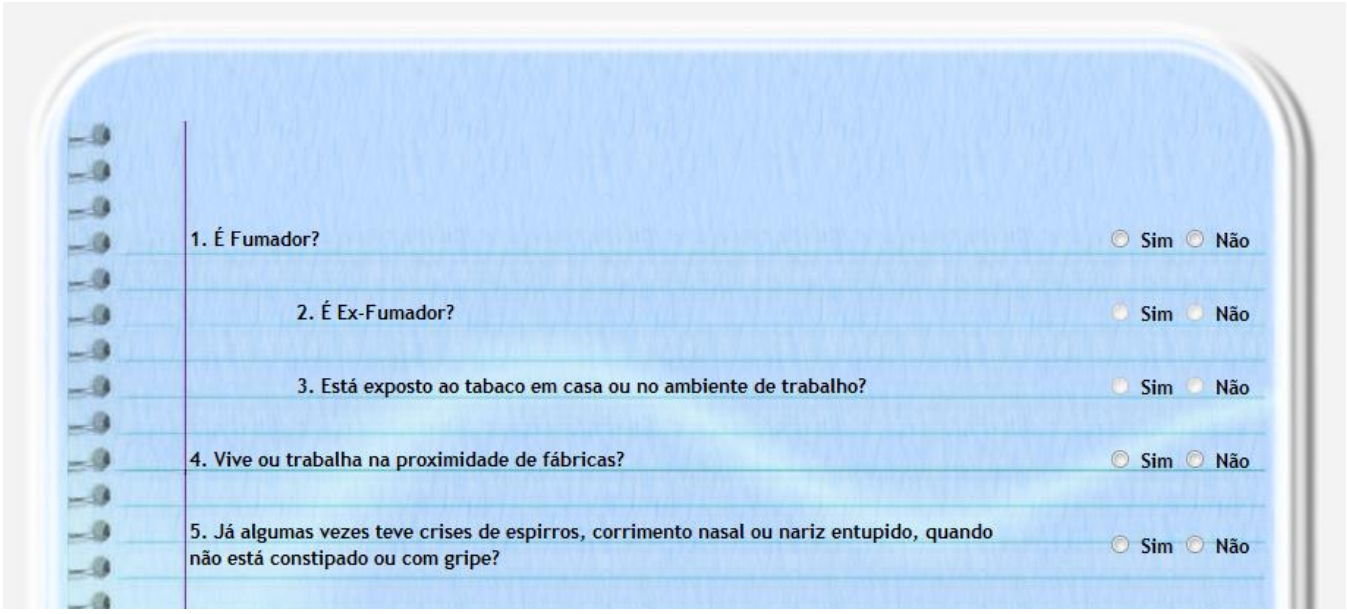


Figura 3.8 - Inquérito (1)

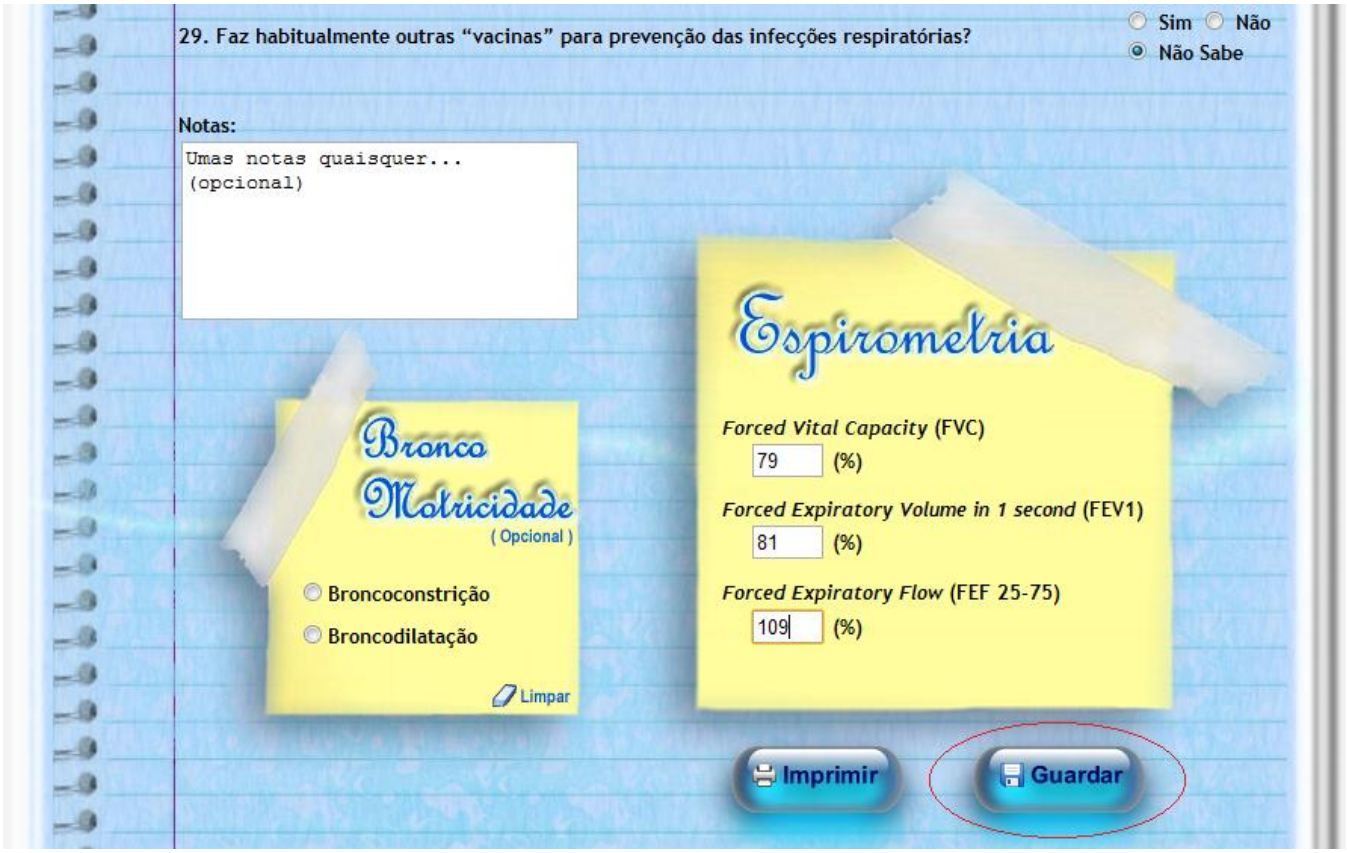


Figura 3.9 - Inquérito (2)

Quando todos os campos estiverem devidamente preenchidos ou seleccionados, o Utilizador tem a opção de imprimir o inquérito (botão "Imprimir") antes de guardar os dados no sistema (botão "Guardar").

Ao carregar no botão "Pesquisar", para verificar a existência do utente no sistema e para a pesquisa ser a mais eficiente possível, é aconselhada a procura do utente ou pelo seu nome ou pela sua data de nascimento, bastando para tal digitar os caracteres no campo reservado para o efeito.

É igualmente possível ordenar os dados pelo "Nome", "Data de Nascimento" e "Último Inquérito" descendentemente ou ascendentemente, bastando para tal clicar no respectivo cabeçalho.

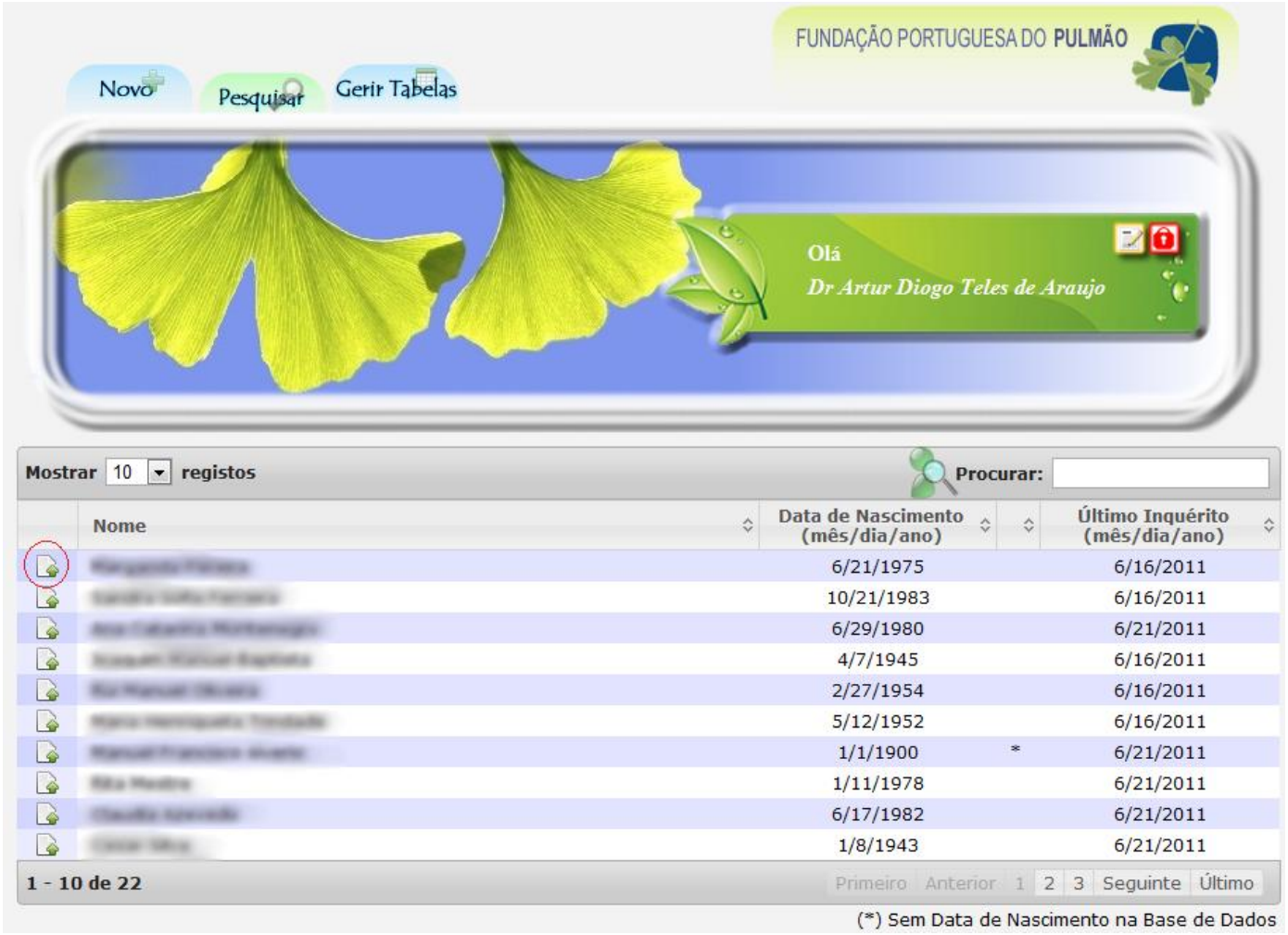


Figura 3.10 - Pesquisar Utente

Quando aparecer o utente desejado na tabela, carregar no botão circundado a vermelho na Figura 3.10 para ter acesso aos dados do seu último inquérito.

Após o passo anterior, deverá aparecer uma página com os dados do utente, do seu inquérito e da espirometria referentes à última vez que participou numa iniciativa.

A partir daqui, será mais fácil dar seguimento ao novo inquérito que vai ser modificado segundo as novas respostas do utente. Isto acontece porque os campos do seu último inquérito já se encontram seleccionados e/ou preenchidos e apenas alguns dados vão mudando com o tempo, sendo outros naturalmente constantes.

Ao seleccionar "Gerir Tabelas" no menu principal, uma imagem semelhante à Figura 3.11 deverá aparecer. Existem três tabelas que podem ser geridas pelo(s) administrador(es) do sistema: "Utentes", "Utilizadores" e "Populações".

Figura 3.11 - Gerir Tabelas

A tabela "Utentes" (Figura 3.12) possibilita a apresentação de todos os dados de qualquer utente registado no sistema ( ), a edição dos seus dados pessoais ( ), assim como a sua eliminação ( ). Ao eliminar um utente, todos os dados dos inquéritos realizados pelo mesmo serão também apagados. Na tabela são mostrados os dados mais relevantes do utente.

É de notar que o funcionamento da tabela (e.g. procura e ordenações) é idêntico ao da tabela apresentada quando se selecciona o menu "Pesquisar".

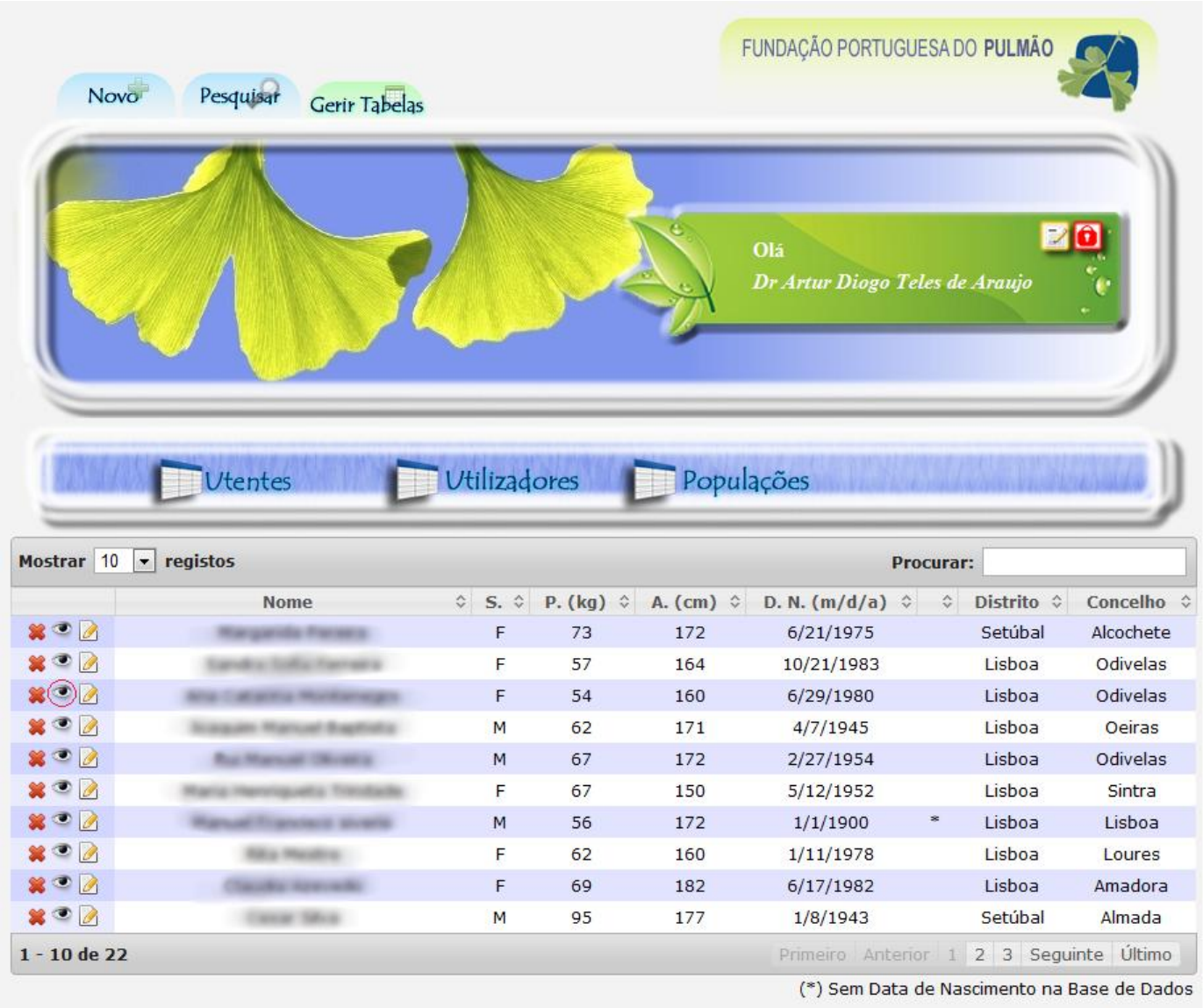


Figura 3.12 - Gerir Tabelas: Utentes

Ao ver todos os detalhes do Utente na nova janela que aparece, clicando na imagem circundada a vermelho (Figura 3.13), todos os inquéritos realizados serão apresentados. Os inquéritos estão ordenados do mais recente ao mais antigo. Para cada inquérito apresentado também é mostrada a data da sua realização assim como o utilizador que o preencheu. Existe também a possibilidade de eliminar o inquérito seleccionado.

Ao clicar na data do inquérito, será o próprio inquérito e as suas respectivas respostas que aparecerão. Por outro lado, ao clicar no nome do utilizador que preencheu o inquérito, são os seus dados que serão mostrados (Figura 3.14).

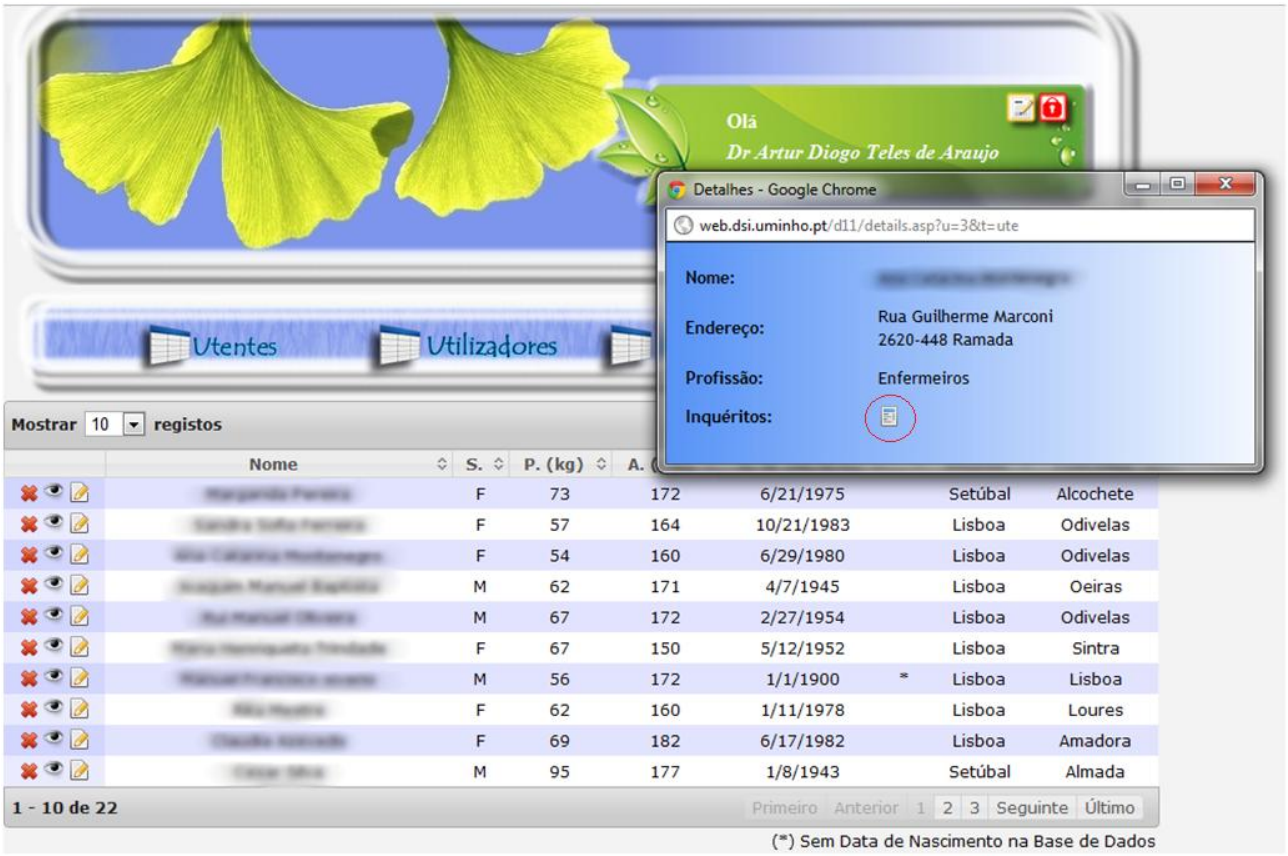


Figura 3.13 - Gerir Tabelas: Detalhes do Utente

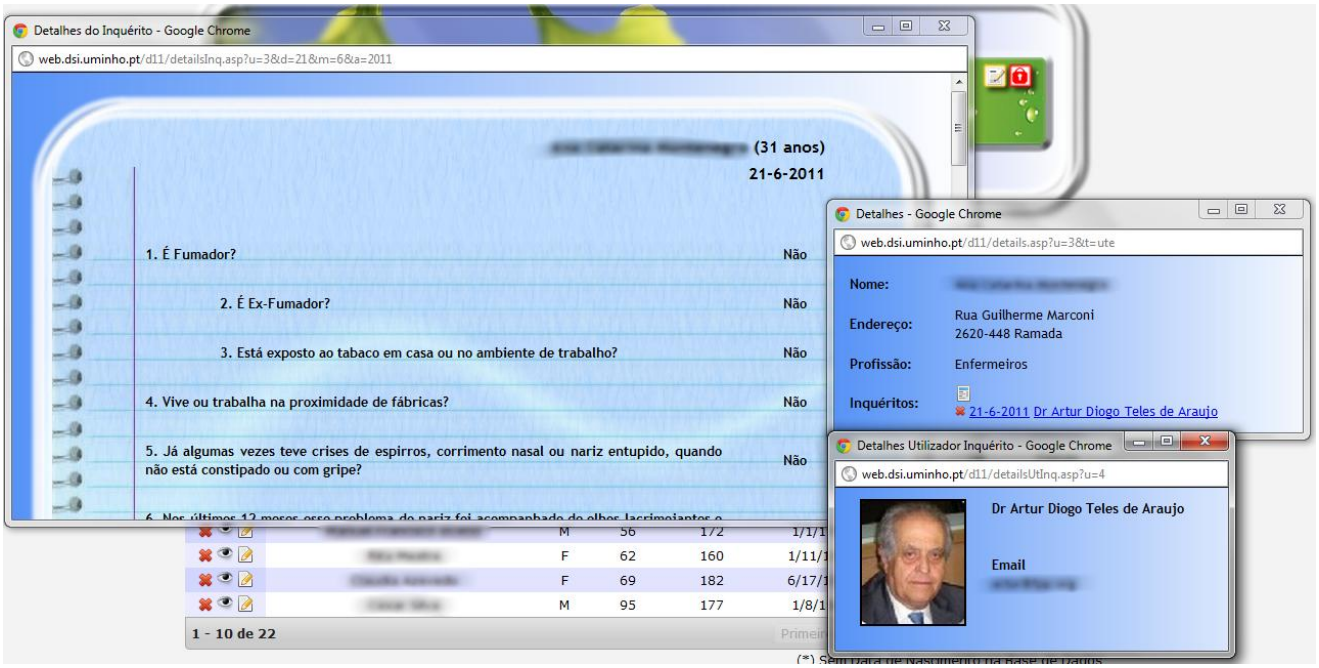


Figura 3.14 - Gerir Tabelas: Detalhes do inquérito do utente e do utilizador que o preencheu

A tabela "Utilizadores" permite a apresentação dos dados principais de todos os utilizadores registados no sistema, assim como a sua eliminação. Na tabela são mostrados os dados mais relevantes do utilizador. Contudo, à semelhança do que acontece com a tabela dos Utentes, ao clicar na imagem representada por um olho, todos os seus outros dados serão apresentados. Para acrescentar um Utilizador ao sistema, basta clicar em "Adicionar Utilizador" (Figura 3.15).

É de notar que o funcionamento da tabela é idêntico ao das outras tabelas apresentadas.

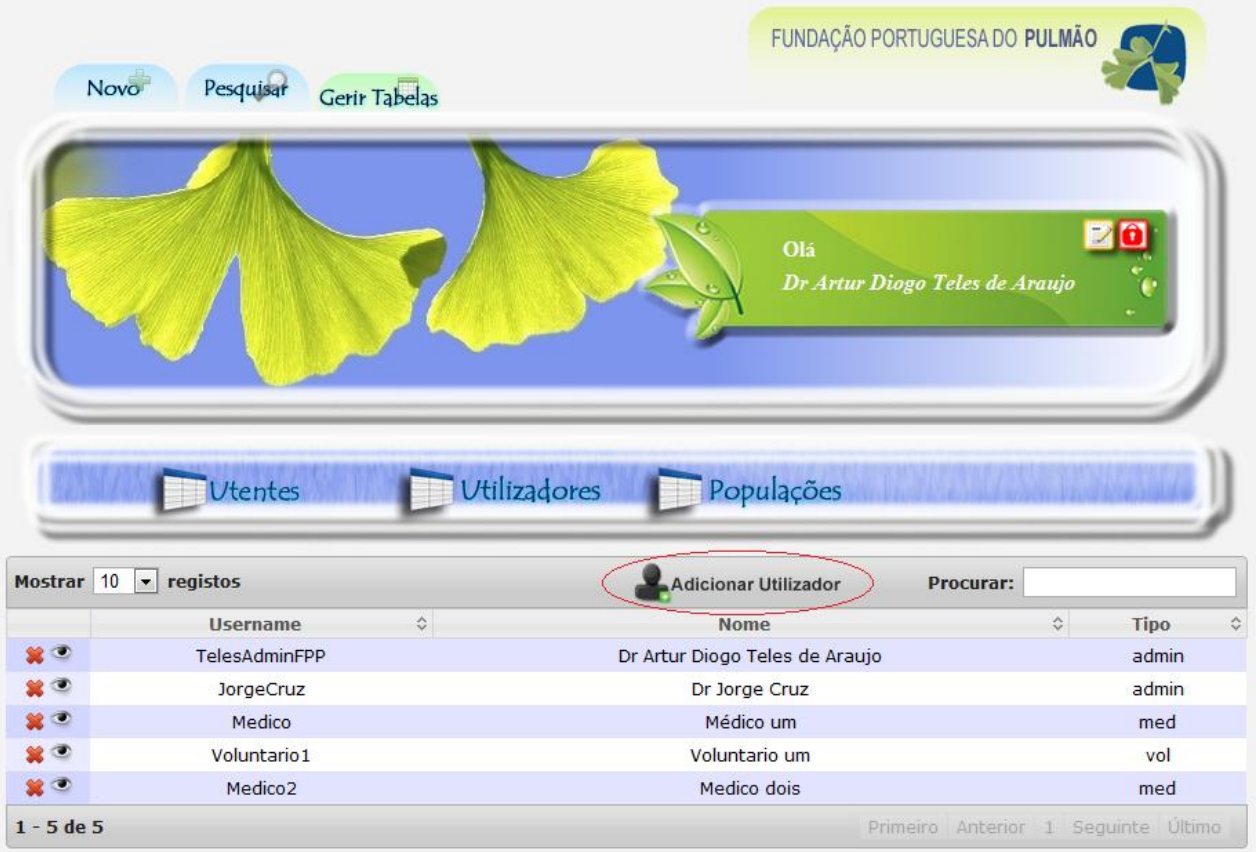


Figura 3.15 - Gerir Tabelas: Utilizadores

Na nova janela que aparece (Figura 3.16), é necessário preencher todos os dados relativos ao novo Utilizador à excepção do seu e-mail. De notar ainda que a *password* tem de conter no mínimo 6 caracteres e no máximo 12. Existem três tipos de utilizadores: Administrador, Médico (que preenche inquéritos e pode ver aqueles que preencheu) e Voluntário (apenas pode preencher os inquéritos).

Figura 3.16 - Gerir Tabelas: Novo Utilizador

A tabela "Populações" (Figura 3.17) possibilita a apresentação de todos os grupos populacionais registados, assim como a criação e eliminação dos mesmos, mas também a visualização dos Utentes pertencentes de uma determinada população (Figura 3.18).

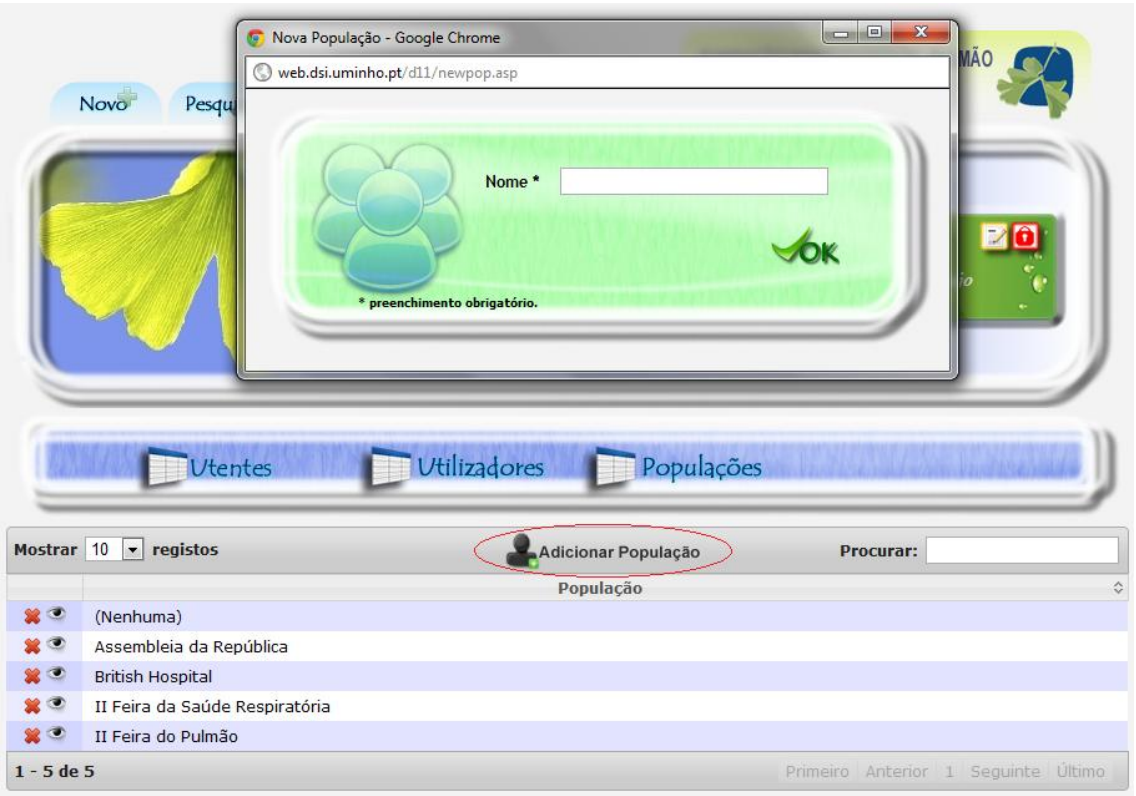


Figura 3.17 - Gerir Tabelas: Populações (Adicionar)

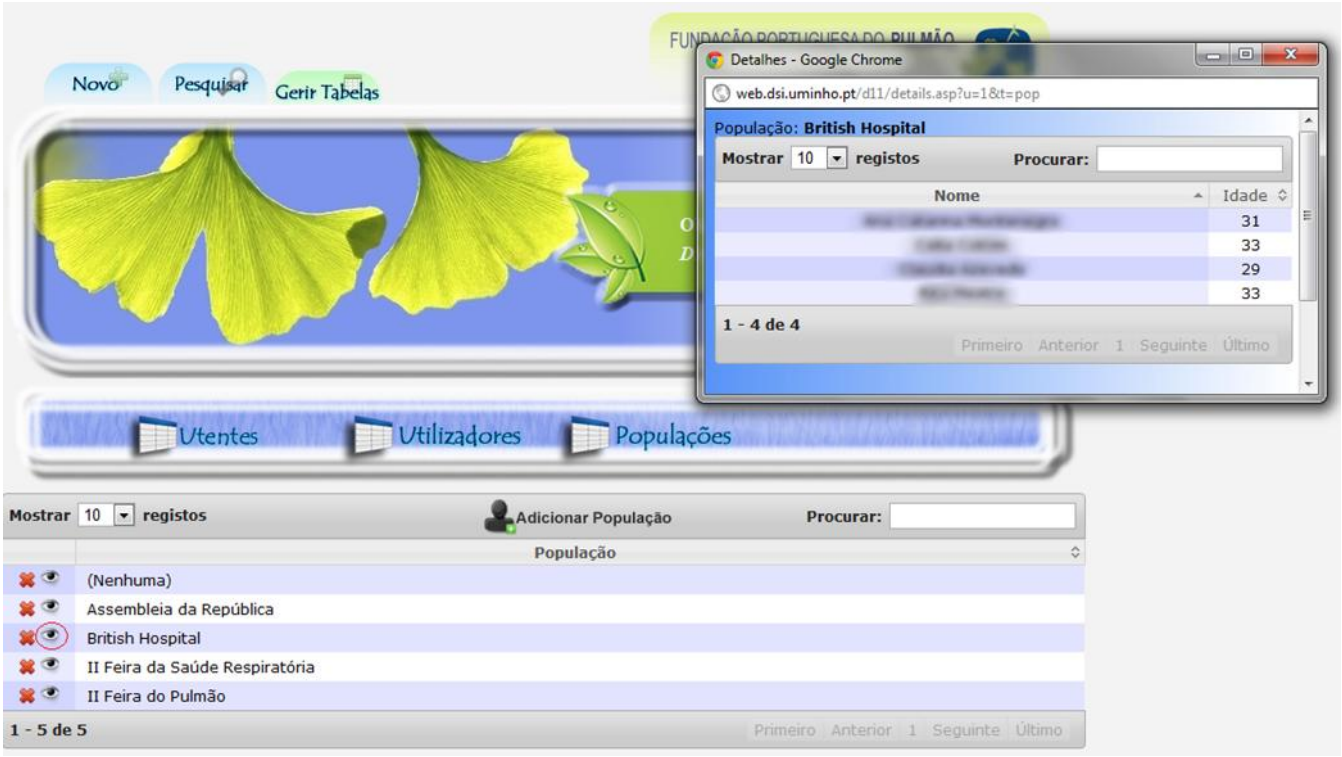


Figura 3.18 - Gerir Tabelas: Populações (Detalhes)

Como referido anteriormente, existem três tipos de utilizadores: Administrador, Médico e Voluntário. O que foi apresentado acima é referente aos menus acedidos com uma autenticação feita por um Administrador. É portanto importante salientar as diferenças existentes entre as três autenticações possíveis. Enquanto que o Administrador pode dar início a um novo inquérito para um novo utente ("Novo"), pesquisar por um utente já existente no sistema, dar seguimento ao seu processo com o último inquérito respondido ("Pesquisar") e ter acesso à gestão das tabelas Utentes, Utilizadores e Populações ("Gerir Tabelas"), os voluntários apenas têm acesso aos menus "Novo" e "Pesquisar". Os médicos, para além de também terem acesso a estes dois menus, podem visualizar os inquéritos realizados por si ("Os meus Inquéritos").

3.3.2 *Data Warehouse*

Depois de vistas algumas das funcionalidades da aplicação *Web*, a modelação do DW, assim como o processo de ETL serão apresentados nesta subsecção.

Em primeiro lugar, é importante referir que a modelação do DW foi definida com base no processo de tomada de decisão da FPP. Este modelo de dados integra um vector de análise, representado pela tabela de factos "FactFPP". A modelação multidimensional deste esquema em estrela é ilustrada pela Figura 3.19. Adoptou-se a língua inglesa na modelação do DW para facilitar a publicação de resultados associados a este trabalho em fóruns internacionais, como conferências, seminários ou workshops.

A tabela de factos "FactFPP" permite o armazenamento da informação relevante recolhida através do questionário. Esta tabela está ligada a várias tabelas dimensões que permitem a análise desses dados sob diversas perspectivas. Olhando para o modelo do DW, pode ser observado que esta tabela de factos está ligada às dimensões "*Time*" (Tempo), "*Location*" (Localização), "*Profession*" (Profissão), "*Patient*" (Utente), "*Smoke Characterization*" (Caracterização do Fumo), "*Allergy Characterization*" (Caracterização das Alergias), "*Cough Characterization*" (Caracterização da Tosse), "*Fatigue Characterization*" (Caracterização do Cansaço) e "*Pulmonary Diseases Characterization*" (Caracterização das Doenças Pulmonares), o que significa que o "FVC", o "FEV1" e o "FEF 25-75" (*Forced Expiratory Flow* 25%–75%), três valores obtidos durante o exame espirométrico que caracteriza a DPOC, e o "*Severity Stage*" (Nível de Gravidade), podem ser analisados recorrendo a questões como "quando?" e "onde?" é que a DPOC se verifica, "quem?" (com a informação dos indivíduos associados como a idade, sexo, peso, entre outros atributos) e "como?" (com as diversas questões do questionário agrupadas em cinco dimensões distintas). O facto "*Patient*" é um contador de eventos utilizado para quantificar o número de indivíduos com sintomas ou características específicas. É de notar ainda que a dimensão "*Profession*" e os atributos assinalados com um * na Figura 3.19 (presentes noutras dimensões e também na tabela de factos) não serão utilizados nestas primeiras análises presentes no relatório de dissertação (tanto OLAP como DM) porque estes dados não estão presentes no conjunto de dados disponibilizado pela FPP. Todavia, como a implementação da aplicação *Web* já permite a recolha desses dados, estes serão incluídos em análises futuras.

O modelo do DW apresentado foi desenhado a pensar na evolução do esquema em estrela para um esquema em constelação, à medida que novas tabelas de factos vão sendo adicionadas. Neste momento, está previsto o acréscimo de duas tabelas de factos: uma para o estudo da pneumonia e outra para o estudo do cancro do pulmão. Com o crescimento da constelação, mais sintomas e dados sobre os indivíduos podem ser relacionados no estudo de uma ou mais doenças.

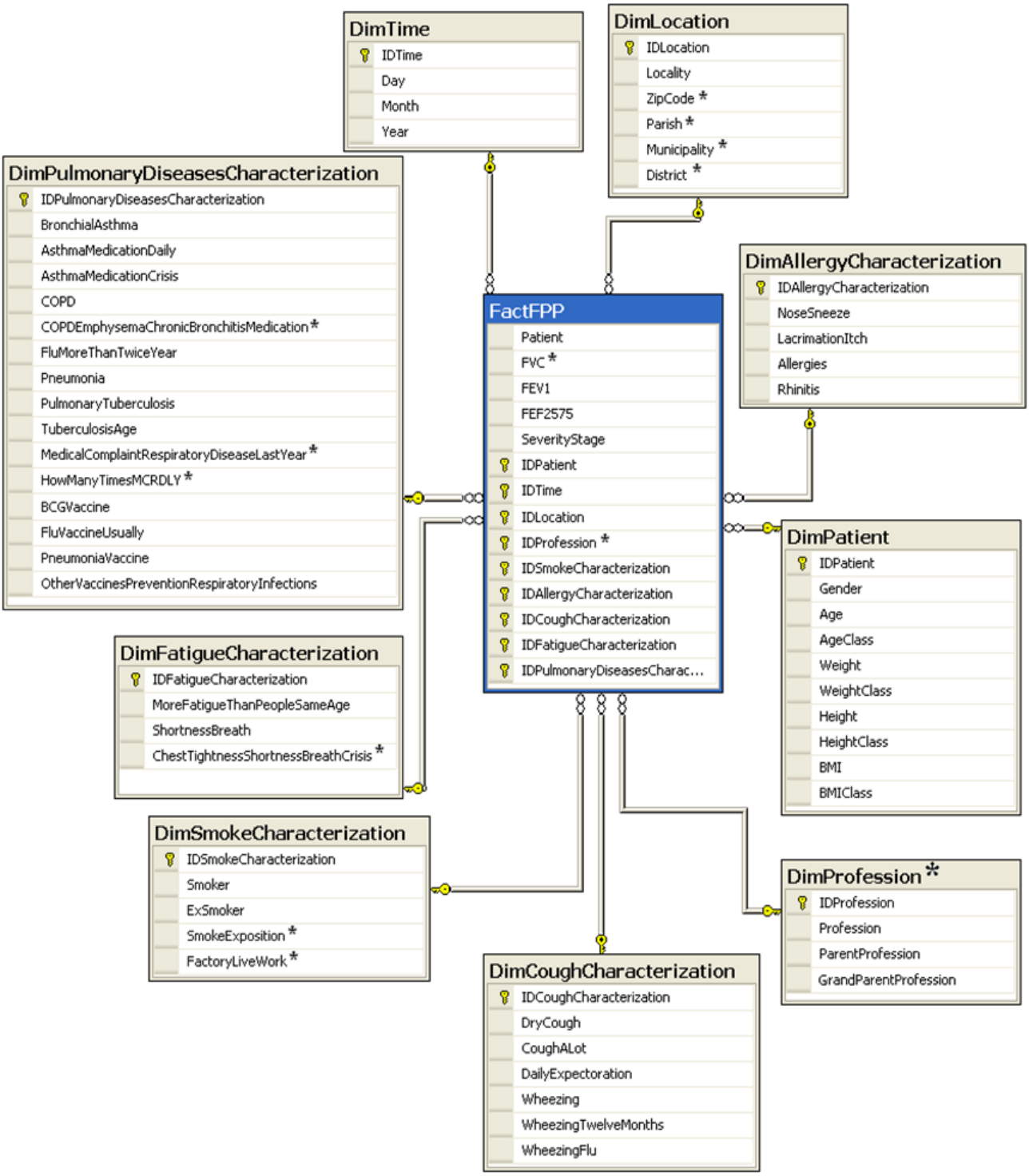


Figura 3.19 - Modelo de Dados do *Data Warehouse*

Depois de apresentado o modelo de dados do DW, de seguida são detalhados os diversos atributos que integram a tabela de factos e as tabelas de dimensão.

- **FactFPP**: Esta tabela é a única tabela de factos do modelo e contém os diferentes factos que podem ser analisados recorrendo às várias dimensões ligadas à própria (Tabela 3.14). De notar que o nível de gravidade da DPOC (*SeverityStage*) é calculado utilizando o FEV1. Se 80 ≤ FEV1 ≤ 50, então *SeverityStage*=2 (Moderado). Se 49 ≤ FEV1 ≤ 30, então *SeverityStage*=3 (Grave). Se 29 ≤ FEV1 ≤ 0, então *SeverityStage*=4 (Muito Grave);

FactFPP				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
Patient	Int			Contador de eventos
FEV1	Int			Valor do FEV1 registado na espirometria
FEF2575	Int			Valor do FEF 25%-75% registado na espirometria
FVC *	Int			Valor do FVC registado na espirometria
SeverityStage	Int			Nível de Gravidade da DPOC
IDPatient	Int	X	X	ID da dimensão Utente
IDTime	Int	X	X	ID da dimensão Tempo
IDLocation	Int	X	X	ID da dimensão Localização
IDProfession *	Int	X	X	ID da dimensão Profissão
IDSmokeCharacterization	Int	X	X	ID da dimensão que caracteriza o fumo
IDAllergyCharacterization	Int	X	X	ID da dimensão que caracteriza as alergias
IDCoughCharacterization	Int	X	X	ID da dimensão que caracteriza a tosse
IDFatigueCharacterization	Int	X	X	ID da dimensão que caracteriza o cansaço
IDPulmonaryDiseasesCharacterization	Int	X	X	ID da dimensão que caracteriza as doenças pulmonares

Tabela 3.14 - Tabela "FactFPP" do DW

- **DimTime**: Esta dimensão regista as datas dos inquéritos efectuados (Tabela 3.15);

DimTime				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDTime	Int	X		ID da dimensão Tempo
Day	Int			Dia do inquérito
Month	Int			Mês do inquérito
Year	Int			Ano do inquérito

Tabela 3.15 - Tabela "DimTime" do DW

- **DimProfession**: Dimensão onde estão presentes os diferentes níveis hierárquicos da profissão do utente (Tabela 3.16);

DimProfession *				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDProfession	Int	X		ID da dimensão Profissão
Profession	nvarchar(max)			Designação da Profissão de Nível 3
ParentProfession	nvarchar(max)			Designação da Profissão de Nível 2
GrandParentProfession	nvarchar(max)			Designação da Profissão de Nível 1

Tabela 3.16 - Tabela "DimProfession" do DW

- **DimLocation**: Esta dimensão regista as localizações das residências dos utentes (Tabela 3.17);

DimLocation				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDLocation	Int	X		ID da dimensão Localização
Locality	nvarchar(50)			Localidade de residência do utente
ZipCode *	nvarchar(8)			Código Postal de residência do utente
Parish *	nvarchar(50)			Freguesia de residência do utente
Municipality *	nvarchar(50)			Concelho de residência do utente
District *	nvarchar(50)			Distrito de residência do utente

Tabela 3.17 - Tabela "DimLocation" do DW

- **DimPatient**: Dimensão onde estão presentes os diferentes dados pessoais do utente (Tabela 3.18). Com as várias interacções com a FPP foram identificadas quatro classes de idade (*AgeClass*): [0-17], [18-40], [41-64] e 65+. Também foram definidas, do mesmo modo, quatro classes de peso (*WeightClass* – [0-50], [51-70], [71-90] e 91+), altura (*HeightClass* – [0-120], [121-160], [161-180], 180+) e Índice de Massa Corporal (*BMIClass* – [0-18.4], [18.5-24.9], [25.0-29.9], 30.0+). O Índice de Massa Corporal é calculado usando a seguinte fórmula:

  $$Peso\ (kg) \div \left(\frac{Altura\ (cm)}{100}\right)^2 .$$

  O intervalo [0-18.4] corresponde a um peso abaixo do peso ideal, [18.5-24.9] a um peso normal, [25.0-29.9] a um peso acima do normal, e 30.0+ corresponde à classificação "Obesidade";

DimPatient				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDPatient	Int	X		ID da dimensão Utente
Gender	nvarchar(1)			Sexo do Utente
Age	Int			Idade do Utente (anos)
AgeClass	nvarchar(50)			Classe da idade do Utente
Weight	Int			Peso do Utente (kg)
WeightClass	nvarchar(50)			Classe do peso do Utente
Height	Int			Altura do Utente (cm)
HeightClass	nvarchar(50)			Classe da altura do Utente
BMI	decimal(4,2)			Índice de Massa Corporal do Utente
BMIClass	nvarchar(50)			Classe do Índice de Massa Corporal do Utente

Tabela 3.18 - Tabela "DimPatient" do DW

As restantes cinco dimensões foram construídas agrupando conjuntos de respostas que pelas suas características devem pertencer à mesma dimensão.

- **DimAllergyCharacterization**: Esta dimensão agrupa as respostas possíveis às perguntas que estão relacionadas com as alergias dos utentes (Tabela 3.19);

DimAllergyCharacterization				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDAllergyCharacterization	Int	X		ID da dimensão que caracteriza as alergias
NoseSneeze	nvarchar(50)			Resposta à pergunta: "Já algumas vezes teve crises de espirros, corrimento nasal ou nariz entupido, quando não está constipado ou com gripe?"
LacrimationItch	nvarchar(50)			Resposta à pergunta: "Nos últimos 12 meses esse problema do nariz foi acompanhado de olhos lacrimejantes e com comichão?"
Allergies	nvarchar(50)			Resposta à pergunta: "Já alguma vez um médico lhe disse sofrer de alergias?"
Rhinitis	nvarchar(50)			Resposta à pergunta: "Já alguma vez um médico lhe disse sofrer de rinite alérgica?"

Tabela 3.19 - Tabela "DimAllergyCharacterization" do DW

- **DimCoughCharacterization**: Dimensão onde estão agrupadas todas as respostas possíveis às perguntas que estão relacionadas com a tosse dos utentes (Tabela 3.20);

DimCoughCharacterization				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDCoughCharacterization	Int	X		ID da dimensão que caracteriza a tosse
DryCough	nvarchar(50)			Resposta à pergunta: "Tem períodos de tosse seca, irritativa?"
CoughALot	nvarchar(50)			Resposta à pergunta: "Tosse muitas vezes na maior parte dos dias? (perguntar se somados todos os dias acha que ultrapassa os três meses)"
DailyExpectoration	nvarchar(50)			Resposta à pergunta: "Tem expectoração ou "catarro" na maioria dos dias?"
Wheezing	nvarchar(50)			Resposta à pergunta: "Já alguma vez teve pieira ou assobios (chieira, apertos no peito)?"
WheezingTwelveMonths	nvarchar(50)			Resposta à pergunta: "Isso aconteceu nos últimos 12 meses?"
WheezingFlu	nvarchar(50)			Resposta à pergunta: "Isso aconteceu quando estava constipado com gripe?"

Tabela 3.20 - Tabela "DimCoughCharacterization" do DW

- **DimSmokeCharacterization**: Esta dimensão guarda todas as respostas associadas às perguntas que estão relacionadas com o fumo (Tabela 3.21);

DimSmokeCharacterization				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDSmokeCharacterization	Int	X		ID da dimensão que caracteriza o fumo
Smoker	nvarchar(50)			Resposta à pergunta: "É fumador?"
ExSmoker	nvarchar(50)			Resposta à pergunta: "É ex-Fumador?"
SmokeExposition *	nvarchar(50)			Resposta à pergunta: "Está exposto ao tabaco em casa ou no ambiente de trabalho?"
FactoryLiveWork *	nvarchar(50)			Resposta à pergunta: "Vive ou trabalha na proximidade de fábricas?"

Tabela 3.21 - Tabela "DimSmokeCharacterization" do DW

- **DimFatigueCharacterization**: Esta dimensão agrupa as respostas possíveis às perguntas que estão relacionadas com o cansaço dos utentes (Tabela 3.22);

DimFatigueCharacterization				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDFatigueCharacterization	Int	X		ID da dimensão que caracteriza o cansaço
MoreFatigueThanPeopleSameAge	nvarchar(50)			Resposta à pergunta: "Cansa-se mais do que as outras pessoas da sua idade?"
ShortnessBreath	nvarchar(50)			Resposta à pergunta: "Tem que parar com "falta de ar" ao subir uma ladeira ou um lanço de escada?"
ChestTightnessShortnessBreathCrisis *	nvarchar(50)			Resposta à pergunta: "Tem por vezes crises súbitas de aperto no peito ou falta de ar?"

Tabela 3.22 - Tabela "DimFatigueCharacterization" do DW

- **DimPulmonaryDiseasesCharacterization**: Dimensão onde estão agrupadas todas as respostas associadas às perguntas que estão relacionadas com as doenças pulmonares (Tabela 3.23);

DimPulmonaryDiseasesCharacterization				
Campo	**Tipo**	**Chave**		**Descrição**
		P	**E**	
IDPulmonaryDiseasesCharacterization	Int	X		ID da dimensão que caracteriza as doenças pulmonares
BronchialAsthma	nvarchar(50)			Resposta à pergunta: "Já alguma vez um médico lhe disse sofrer de asma brônquica?"
AsthmaMedicationDaily	nvarchar(50)			Resposta à pergunta: "Faz medicação para a asma diariamente?"
AsthmaMedicationCrisis	nvarchar(50)			Resposta à pergunta: "Faz medicação para a asma (só em crises)?"
COPD	nvarchar(50)			Resposta à pergunta: "Já alguma vez um médico lhe disse sofrer de bronquite crónica, DPOC (Doença Pulmonar Obstrutiva Crónica) ou enfisema?"
COPDEmphysemaChronicBronchitisMedication	nvarchar(50)			Resposta à pergunta: "Faz medicação para a bronquite crónica, DPOC ou enfisema?"
FluMoreThanTwiceYear	nvarchar(50)			Resposta à pergunta: "Constipa-se ou tem gripe mais do que duas vezes no ano?"
Pneumonia	nvarchar(50)			Resposta à pergunta: "Já teve pneumonia ou broncopneumonia?"
PulmonaryTuberculosis	nvarchar(50)			Resposta à pergunta: "Já teve tuberculose pulmonar?"
TuberculosisAge	Int			Resposta à pergunta: "Quando?"
MedicalComplaintRespiratoryDiseaseLastYear *	nvarchar(50)			Resposta à pergunta: "No último ano foi ao médico por queixas de doenças respiratórias?"
HowManyTimesMCRDLY *	Int			Resposta à pergunta: "Quantas vezes?"
BCGVaccine	nvarchar(50)			Resposta à pergunta: "Está vacinado com a BCG (vacina contra a Tuberculose)?"
FluVaccine	nvarchar(50)			Resposta à pergunta: "Vacina-se habitualmente contra a gripe?"
PneumoniaVaccine	nvarchar(50)			Resposta à pergunta: "Está vacinado contra a pneumonia?"
OtherVaccinesPreventionRespiratoryInfections	nvarchar(50)			Resposta à pergunta: "Faz habitualmente outras "vacinas" para prevenção das infecções respiratórias?"

Tabela 3.23 - Tabela "DimPulmonaryDiseasesCharacterization" do DW

Após a caracterização dos dados disponíveis, da descrição da implementação realizada para aplicação *Web*, e da modelação do DW tendo sempre em mente o processo de tomada de decisão, o próximo passo está associado com a análise do conjunto de dados que foi inicialmente fornecido pela FPP e realizar as transformações necessárias para o posterior carregamento dos dados para o DW. O processo de ETL começou pela identificação de erros no conjunto de dados inicial e pela sua respectiva correcção, assim como pela homogeneização dos dados. De seguida são resumidos, na Tabela 3.24, Tabela 3.25 e Tabela 3.26, as principais correcções e transformações realizadas ao conjunto de dados inicial.

Designação Atributo	Valores Registados	Descrição e valores possíveis	Dados em Falta / Erros nos dados e Correcções Efectuadas
Idade	[4, 88]	Idade do utente (número inteiro)	
Sexo	{m, f, n}	Sexo do utente {m, f}	n (1 registo) -> corrigir para m
Localidade	(localidades)	Descrição da localidade (nome de localidade)	Respostas não padronizadas (ex: "Lisboa", "9700", em branco). Exemplos de correcções: SD Rana -> São Domingos de Rana, 2610->Amadora, em branco->Desconhecido
Peso	[6, 130]	Peso do utente (número inteiro)	49 registos sem indicação do peso Moda: 70 / Média: 68 Substituiu-se os registos em branco pela Moda.
Altura	[58, 197], 1157, Null	Altura do utente (número inteiro)	1157->157 (dada a idade, 26 anos, e o peso, 53 Kg, o valor correcto deve ser 157cm) 24 registos sem indicação da altura Moda: 160 / Média: 165 Substituiu-se os registos em branco pela Moda.
IMC	[0, 65.2], Null	Índice de Massa Corporal do utente (número)	Diversos valores errados 70 registos vazios (já contanto com os que têm o valor 0) Depois de corrigidos os valores do peso e da altura, os IMC foram calculados automaticamente.
Tabaco	{s, n, ex, m}	Resposta às perguntas do inquérito "É fumador? É ex-Fumador?" {"s": sim, "n": não, "ex": ex-fumador}	1 registo com o valor m -> corrigido para n
Nariz / Espirros	{a, ex, n, s}	Resposta à pergunta do inquérito "Já algumas vezes teve crises de espirros, corrimento nasal ou nariz entupido, quando não está constipado ou com gripe?" {s, n}	1 registo com o valor a -> corrigido para s 1 registo com o valor ex -> troca de valores com a coluna Tabaco
Lacrimejo / Comichão	{n, s}	Resposta à pergunta do inquérito "Nos últimos 12 meses esse problema do nariz foi acompanhado de olhos lacrimejantes e com comichão?" {s, n}	

Tabela 3.24 - Tabela Resumo da Compreensão dos Dados e Detecção de Anomalias (1)

Tosse seca	{n, s}	Resposta à pergunta do inquérito "Tem períodos de tosse seca, irritativa (alérgica)?" {s, n}	
Tosse +3 meses	{n, s}, Null	Resposta à pergunta do inquérito "Tosse muitas vezes na maior parte dos dias? (se somados todos os dias acha que ultrapassa os três meses?)" {s, n}	1 registo a Null -> corrigido para "n" já que a coluna Tosse seca apresenta também o valor "n"
Expectoração diária	{n, s}	Resposta à pergunta do inquérito "Tem expectoração ou "catarro" na maioria dos dias?" {s, n}	
Pieira	{n, s, sn, ss}, Null	Resposta à pergunta do inquérito "Já alguma vez teve pieira ou assobios (chieira, apertos no peito)?" {s, n}	1 registo com o valor ss -> substituído por "s" (analisando a coluna Pieira últimos 12 meses) 1 registo com o valor sn -> substituído por "s" (analisando a coluna Pieira últimos 12 meses) 1 registo com o valor Null -> substituído por "s" (analisando a coluna Pieira últimos 12 meses)
Pieira últimos 12 meses	{n, s}	Resposta à pergunta do inquérito "Isso aconteceu nos últimos 12 meses?(pieira)" {s, n}	
Pieira com gripe	{n, s}	Resposta à pergunta do inquérito "Isso aconteceu quando estava constipado com gripe? (pieira)" {s, n}	
Cansaço	{n, s}	Resposta à pergunta do inquérito "Cansa-se mais do que as outras pessoas da sua idade?" {s, n}	
Falta de ar	{n, s}	Resposta à pergunta do inquérito "Tem que parar com "falta de ar" ao subir uma ladeira ou um lanço de escada?" {s, n}	
Alergias	{n, s}	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de alergias?" {s, n}	
Rinite	{n, ns, s}, Null	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de rinite alérgica?" {s, n}	1 registo com o valor ns -> manteve-se "ns" (ns - não sabe) visto que é essa a informação disponível na FPP 1 registo com o valor Null -> substitui-se por "ns"
Asma	{n, s}	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de asma brônquica?" {s, n}	
Medicação asma diária	{n, s}, Null	Resposta à pergunta do inquérito "Faz medicação para a asma? (diariamente)" {s, n}	1 registo com o valor Null -> substituído por "n" já que o indivíduo não tem asma
Medicação asma crises	{n, s}	Resposta à pergunta do inquérito "Faz medicação para a asma? (só em crises)" {s, n}	
DPOC	{n, s}	Resposta à pergunta do inquérito "Já alguma vez um médico lhe disse sofrer de bronquite crónica, DPOC (Doença Pulmonar Obstrutiva Crónica) ou enfisema?" {s, n}	

Tabela 3.25 - Tabela Resumo da Compreensão dos Dados e Detecção de Anomalias (2)

Gripe +2xano	{n, ns, s}	Resposta à pergunta do inquérito "Constipa-se ou tem gripe mais do que duas vezes no ano?" {s, n}	2 registos com o valor ns -> manteve-se "ns" visto que é essa a informação disponível na FPP
Pneumonia	{n, ns, s, ss}	Resposta à pergunta do inquérito "Já teve pneumonia ou broncopneumonia?" {s, n, ns}	1 registo com o valor ss -> substituiu-se por "s"
TB	{n, ns, s, s-Ano, s-Idade}	Resposta à pergunta do inquérito "Já teve tuberculose pulmonar? Quando?" {s, n, ns} (Idade[0-120])	Respostas não padronizadas (ex: "1940", "aos 50 anos", "+ de 40 anos") Foi necessário tratar os casos em que apresenta o "quando" (nova coluna idade -> calculou-se a idade para todos os indivíduos). os valores que ficavam a branco foram substituídos por "-1"
BCG	{n, ns, s}	Resposta à pergunta do inquérito "Está vacinado com a BCQ (vacina contra a Tuberculose)?" {s, n, ns}	
Vacina Gripe	{n, ns, s}	Resposta à pergunta do inquérito "Vacina-se habitualmente contra a gripe?" {s, n, ns}	
Vacina Pneumonia	{n, ns, s}	Resposta à pergunta do inquérito "Está vacinado contra a pneumonia?" {s, n, ns}	
Outras Vacinas	{n, ns, s}	Resposta à pergunta do inquérito "Faz habitualmente outras "vacinas" para prevenção das infecções respiratórias?" {s, n, ns}	
FEV1	[14-209], {-, Null}	Valor percentual do FEV1	Valores em falta substituídos por "-1"
FEF 25-75	[9-199], {-, Null}	Valor percentual do FEF 25-75	Valores em falta substituídos por "-1"

Tabela 3.26 - Tabela Resumo da Compreensão dos Dados e Detecção de Anomalias (3)

Depois da limpeza e transformação do conjunto de dados inicial (correcções de inconsistências, erros e valores em falta), foi necessário proceder à selecção dos dados a serem importados para as diversas dimensões e para a tabela de factos do DW. Como o conjunto de dados inicial estava numa folha do *Microsoft Office Excel*, esta ferramenta também foi utilizada, juntamente com o *Microsoft SQL Server 2008*, como DSA para proceder às selecções e possíveis transformações adicionais dos dados na sua passagem para as respectivas tabelas do DW. É de salientar ainda que todo este processo de ETL foi realizado utilizando várias *queries* SQL para concretização das operações necessárias (umas naturalmente mais complexas do que outras). Contudo, e para automatizar o processo, foi criada uma *script* em linguagem *Javascript* para a inserção dos dados na tabela de factos. O *script* está disponível em anexo (Anexo A: *Script* para inserção automática de dados na tabela de factos).

É de realçar ainda que pelo facto de que este processo de passagem e transformação dos dados iniciais da folha de cálculo *Microsoft Office Excel* para o DW criado com a ferramenta *Microsoft SQL Server 2008* apenas se iria realizar uma única vez e de que não necessita portanto de um fluxo de tarefas automatizadas para tal, não foi utilizada neste projecto de dissertação a ferramenta que a *Microsoft* dispõe para esse efeito. Todavia, para as futuras análises dos dados que serão armazenados na BDO da aplicação *Web*, será vantajoso utilizar a ferramenta *Integration Services* disponível no *Business Intelligence Development Studio* do *Microsoft SQL Server 2008*. De facto, esta ferramenta permitirá criar processos automatizados, controlando o fluxo de tarefas necessárias para todos os passos de ETL requeridos. Este processo criado permitirá a extracção, limpeza, transformação e carregamento dos dados da BDO para o DW, mas principalmente o seu refrescamento sem qualquer esforço adicional.

Portanto, após os dados iniciais (presentes numa única tabela) serem extraídos, transformados e carregados para o DW (nove dimensões e uma tabela de factos), estes ficam disponíveis para as posteriores análises, recorrendo a ferramentas OLAP e a algoritmos de DM.

Capítulo 4 – Estudo da incidência da Doença Pulmonar Obstrutiva Crónica

Neste capítulo é descrito o estudo feito sobre os dados presentes no DW. Primeiramente, são justificadas quais foram as selecções e transformações aos dados que foram realizadas para ambas as análises no seu respectivo subcapítulo. São feitas análises aos dados disponíveis, primeiramente recorrendo à componente OLAP, e depois à componente de DM. Os resultados do estudo da incidência da DPOC são evidenciados ao longo do capítulo, para cada uma das abordagens.

4.1 Análise dos dados recorrendo à componente *On-Line Analytical Processing*

Depois de apresentado o modelo do DW que armazena os dados disponíveis para o estudo da DPOC, esta secção apresenta os resultados obtidos utilizando a tecnologia OLAP. Esta tecnologia é usada para analisar os diversos indicadores da tabela de factos sob diversas perspectivas.

Na análise aos dados recorrendo à tecnologia OLAP, foi utilizado um subconjunto dos dados disponíveis. Apenas os registos cujos utentes apresentaram um valor do FEV1 menor do que 80% foram considerados, perfazendo um total de 275 registos. Esta opção prendeu-se com facto de que, no âmbito desta dissertação, interessa analisar apenas os utentes que apresentam DPOC. Esta selecção de utentes que apresentam um valor do FEV1 menor do que 80% permite a caracterização dos sintomas e dos factores mais pessoais (como sexo ou idade) dos indivíduos com DPOC.

Das diversas análises efectuadas, são apresentadas as sete que revelaram ter uma maior importância ao nível da interpretação dos dados disponíveis, ou seja, poderem ser retiradas conclusões mais relevantes e que contribuíram assim para um maior conhecimento dos dados.

A primeira análise apresentada verifica os níveis de gravidade da DPOC dos utentes. A Figura 4.1 mostra que quase todos os utentes, 254 (92.7%), estão no 2º nível de gravidade da DPOC (Moderado). Apenas 18 indivíduos estão no 3º nível de gravidade da doença (Grave) e 3 utentes estão no 4º nível de gravidade da DPOC (Muito Grave).

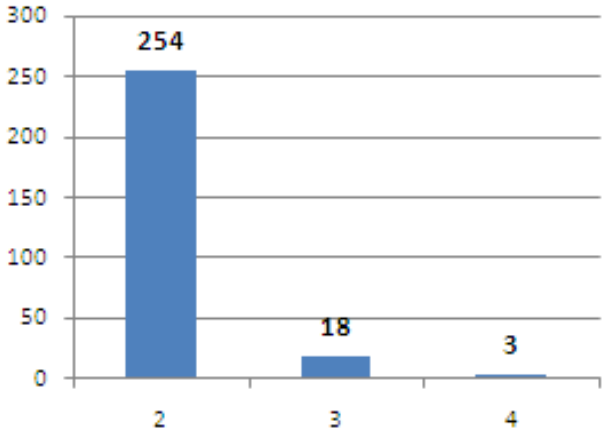


Figura 4.1 - Níveis de Gravidade

Depois de analisar a incidência dos níveis de gravidade da DPOC dos utentes, a próxima análise está focada nas respostas dadas pelos utentes às perguntas que possibilitam a sua caracterização. Como mencionado anteriormente, foi diagnosticada DPOC a todos esses indivíduos presentes no conjunto de dados analisado, depois de realizado o exame espirométrico.

A caracterização obtida pelo grupo de questões relacionadas com o fumo pode ser observada na Figura 4.2. Os resultados mostram que, apesar de o tabaco ser um dos factores de risco mais óbvios para esta doença, nos dados analisados, 168 utentes (61.1%) com diagnóstico de DPOC revelaram nunca ter fumado.

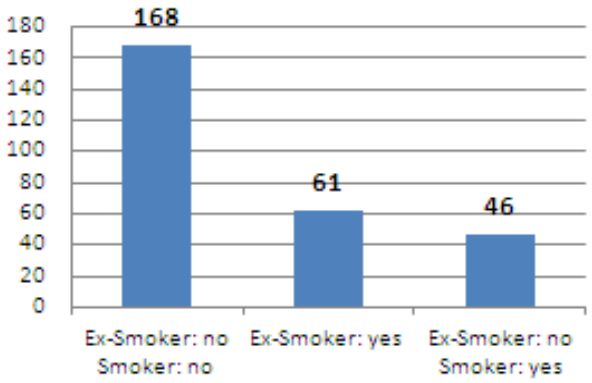


Figura 4.2 - Caracterização das perguntas relacionadas com o fumo

No que diz respeito à caracterização do grupo de perguntas relacionadas com o cansaço dos utentes, os resultados são apresentados na Figura 4.3. Apesar de ser claro que a maioria dos utentes, 187 (68.0%), sente maior cansaço do que as pessoas que têm sensivelmente a mesma idade (MFTPSA - *More Fatigue Than People Same Age*) e/ou tem falta de ar (SB - *Shortness of Breath*) e 118 (42.9%) até sentem os dois sintomas, existem 88 indivíduos (32.0%) da amostra de dados que têm DPOC mas que não apresentam nenhum destes dois sintomas.

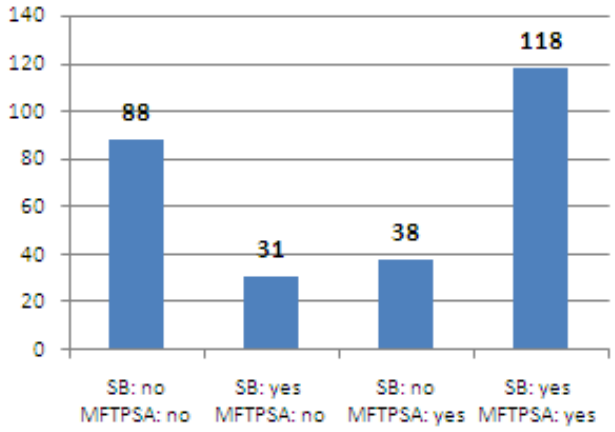


Figura 4.3 - Caracterização das perguntas relacionadas com o cansaço

Analisando as respostas obtidas ao grupo de questões relacionadas com a tosse (Figura 4.4) e mais precisamente os dois sintomas da DPOC presentes neste grupo de perguntas, tosse (*Dry Cough*) e expectoração (DE – *Daily Expectoration*), pode ser verificado que 73 (26.5%) utentes com DPOC têm tosse seca e expectoração diária.

É de referir ainda que, olhando para os dados analisados, estes dois sintomas parecem estar relacionados. De facto, quando os indivíduos não têm tosse seca, apenas 17.3% (18/104) têm expectoração diária. Todavia, quando os utentes têm tosse seca, a percentagem dos mesmos que também tem expectoração diária aumenta consideravelmente, atingindo os 42.7% (73/171).

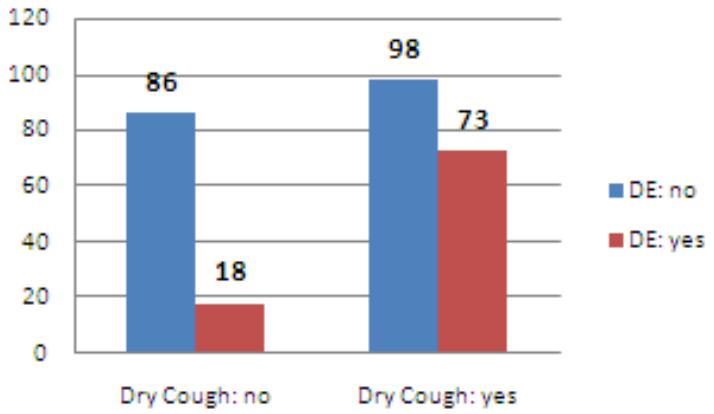


Figura 4.4 - Caracterização das perguntas relacionadas com a tosse

No que concerne ao grupo de perguntas relacionadas com as doenças pulmonares, foram analisadas as três doenças pulmonares presentes neste grupo de perguntas que poderiam ter alguma relação com a DPOC. Essas doenças são as seguintes: Asma brônquica (*Bronchial Asthma*), que também é um dos factores de risco da DPOC; Pneumonia; e Tuberculose Pulmonar (Pulm Tuber – *Pulmonary Tuberculosis*).

Analisando o cubo ilustrado pela Figura 4.5, nenhum dos utentes que apresenta DPOC sofre das três outras doenças em simultâneo. Existem casos de "não sabe" (dk – *do not know*) que não foram considerados nesta observação.

Outra análise que pode ser realçada da Figura 4.5 é que apenas uma pequena percentagem destes utentes que têm DPOC sofrem de pelo menos uma dessas outras três doenças. De facto, 65 (23.6%) indivíduos têm ou tiveram asma brônquica, 57 (20.7%) indivíduos sofrem ou sofreram de pneumonia e 17 (6.2%) de tuberculose pulmonar. Apesar de a asma brônquica apresentar a maior percentagem de incidência entre as três doenças, e como esta doença é um factor de risco da DPOC, era esperado que a sua percentagem de incidência fosse significativamente maior em utentes que têm DPOC.

		Bronchial Asthma		
		no	yes	Grand Total
Pneumonia	Pulm Tuber	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ dk	dk	3	2	5
	no	5	2	7
	Total	8	4	12
⊟ no	dk	2	2	4
	no	155	35	190
	yes	9	3	12
	Total	166	40	206
⊟ yes	dk	6	1	7
	no	25	20	45
	yes	5		5
	Total	36	21	57
Grand Total		210	65	275

Figura 4.5 - Caracterização das perguntas relacionadas com as doenças pulmonares

Observando a caracterização das respostas obtidas ao grupo de perguntas relacionadas com as alergias, os resultados são apresentados na Figura 4.6. Foi analisada a incidência de três características deste grupo de perguntas: olhos lacrimejantes e comichão (LI – *Lacrimation and Itch*); corrimento nasal e espirros (NS – *runny Nose and Sneeze*); e rinite (*rhinithis*).

Os resultados mostram que 202 (73.5%) utentes com DPOC também têm corrimento nasal e espirros quando estes não têm gripe. Este facto pode revelar uma ligação entre estes dois sintomas e a DPOC. Porém, existe também quase o mesmo número de utentes, 201 (73.1%), que afirmam que não sofrem de rinite. Outro facto observado neste grupo de questões foi que, dos 73 utentes com DPOC que não têm corrimento nasal e espirros, 10 (13.7%) têm olhos lacrimejantes e comichão. Dos 202 utentes com DPOC que têm corrimento nasal e espirros, 128 (63.4%) também têm olhos lacrimejantes e comichão.

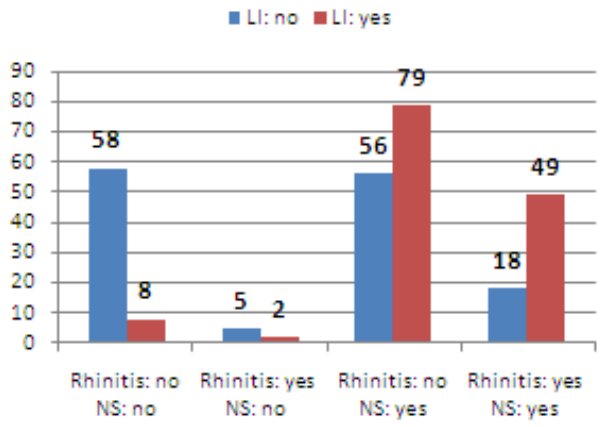


Figura 4.6 - Caracterização das perguntas relacionadas com as alergias

Esta última análise, realizada recorrendo à tecnologia OLAP, tem como objectivo evidenciar a importância das iniciativas e campanhas realizadas a favor desta doença respiratória e a necessidade de participação por parte dos cidadãos. A Figura 4.7 mostra os resultados obtidos. Neste estudo, os utentes foram subdivididos por sexo (*Gender*) e por classes de idade (*Age Class*). O atributo "COPD" indica se o utente já sabia se sofria de DPOC antes de realizar o exame espirométrico nas iniciativas da FPP.

Baseado no conjunto de dados disponível, pode ser observado que 229 utentes (83.3%) com DPOC não sabiam que sofriam desta doença antes de participarem nestas iniciativas promovidas pela FPP. Este cubo confirma também que a idade é um factor de risco da DPOC. De facto, 204 indivíduos (74.2%) que têm esta doença pulmonar têm 41 anos ou mais. É de notar ainda que apenas 7 (2.5%) utentes com menos de 18 anos apresentam DPOC.

		COPD		
		no	yes	Grand Total
Gender	**Age Class**	Fact FPP Count	Fact FPP Count	Fact FPP Count
⊟ f	[0-17]	3		3
	[18-40]	37	2	39
	[41-64]	54	8	62
	65+	44	9	53
	Total	138	19	157
⊟ m	[0-17]	4		4
	[18-40]	22	3	25
	[41-64]	29	4	33
	65+	36	20	56
	Total	91	27	118
Grand Total		229	46	275

Figura 4.7 - Identificação da DPOC nas iniciativas da FPP

4.2 Análise dos dados recorrendo à componente de *Data Mining*

O DM não consegue olhar para o futuro e prever eventos. Todavia, em vez disso, analisa matematicamente o que ocorreu no passado e determina o que é mais provável de ocorrer se a tendência actual continuar, existindo sempre a possibilidade de algum acontecimento alterar esta tendência. Com o DM é possível determinar, com uma confiança determinada estatisticamente, quais são as tendências actuais. As decisões podem, a partir daqui, ser tomadas com base nas tendências identificadas.

Para a análise dos dados disponíveis recorrendo à componente de DM foi seguida a metodologia CRISP-DM.

Como os dois primeiros passos desta metodologia ("Compreensão do Negócio" e "Compreensão dos Dados") já foram executados previamente, sendo necessários para a proposta do modelo do DW, nesta tarefa de análise dos dados com DM apenas faz sentido ser apresentada a partir do passo "Preparação dos Dados". Ao contrário do que aconteceu na análise aos dados recorrendo à tecnologia OLAP, foram utilizados todos os dados disponíveis e não apenas os registos cujos utentes apresentaram um valor do FEV1 menor do que 80%.

A tabela final, que foi a base para as análises efectuadas pelos algoritmos de DM, tem como origem o DW concebido e implementado e resultou de vários *joins* entre a tabela de factos e as várias dimensões do DW (Anexo B: *Query* de criação da tabela principal para análise com *Data Mining*). Foram considerados todos os atributos presentes nas várias dimensões do DW e apenas três atributos da tabela de factos (FEV1, FEF2575 e *SeverityStage*), podendo estes atributos ser posteriormente incluídos, ou não, nos modelos. Como algumas técnicas de modelação têm requisitos específicos, como a análise de determinado tipo de dados, pode ser necessário retroceder à fase de preparação dos dados para convertê-los num determinado formato. Este retrocesso à fase de preparação dos dados também pode acontecer dependendo dos objectivos de DM.

À semelhança do que também aconteceu com a análise dos dados com a componente OLAP, não serão apresentados todos os modelos identificados, mas sim aqueles que revelaram ter uma maior importância ao nível da interpretação dos dados disponíveis, ou seja, poderem ser retiradas conclusões mais relevantes. Espera-se que estes modelos contribuam para um maior conhecimento dos dados e que apresentem previsões e tendências, tentando utilizar e variar ao máximo os algoritmos usados nas técnicas de DM.

Obviamente, a variação dos algoritmos que foram utilizados teve como base as tarefas a realizar. Para um problema de classificação, prever quem tem DPOC atendendo aos sintomas e factores de risco, foram utilizadas as seguintes técnicas de DM: *Microsoft Decision Trees*, *Microsoft Naïve Bayes* e *Microsoft Neural Network*. Para um problema de segmentação, descobrir os padrões escondidos nos dados dos utentes que apresentam DPOC, foi usada a técnica de DM seguinte: *Microsoft Clustering*.

Para prever quem tem DPOC, atendendo aos sintomas e factores de risco (problema de classificação), foram inicialmente considerados todos os dados disponíveis, sendo apenas necessário acrescentar um atributo à tabela: *stringCOPD*. Este atributo indica se o indivíduo sofre de DPOC ("yes") ou não ("no"), baseando-se no atributo *SeverityStage* (que indica o nível de gravidade de DPOC).

A primeira técnica de DM escolhida foi o *Microsoft Decision Trees.* Dos atributos presentes na tabela, não foram considerados a Idade (*Age*), o peso (*Weight*), a altura (*Height*), e o IMC (*BMI*), preferindo utilizar os atributos discretos, agrupados por classe, correspondente de cada um. Também não foram considerados o FEV1, o FEF25-75 e o nível de gravidade de DPOC (*SeverityStage*) pelas seguintes razões: (1) São o FEV1 (principalmente) e o FEF25-75 que definem se um indivíduo apresenta ou não DPOC; (2) o *SeverityStage* é calculado pelo FEV1, isto é, por exemplo, se um indivíduo não tem DPOC o seu *SeverityStage* terá sempre o valor 0. Se estes últimos atributos fossem incluídos nos modelos, estes influenciariam os resultados dos algoritmos. É de realçar ainda que, das várias divisões efectuadas aos conjuntos de dados para treino e para teste (e consequentes testes dos algoritmos para cada hipótese testada), a segmentação que obteve melhores resultados foi a seguinte: 70% dos dados para treino e, por consequência, 30% dos dados para teste.

A Figura 4.8 mostra quais foram os atributos que foram considerados como input e como atributo de previsão. Foram testados os três métodos que calculam o *score* de divisão da árvore (*Entropy*, *Bayesian with K2 Prior* e *Bayesian Dirichlet Equivalent with Uniform prior*). O método que mostrou melhores resultados foi o *Shannon's entropy*. Foi ainda alterado outro parâmetro da árvore, limitando o número mínimo de casos que uma folha deve conter a 50 do conjunto total de treino (1280 utentes). Esta foi uma decisão tomada para apresentar as folhas mais relevantes e, de uma certa forma, podar uma árvore que apresentava nove níveis de nodos (muitos deles pouco relevantes e com muito poucos dados).

Structure	Test Table All_DT1
	Microsoft_Decision_Trees
Age Class	Input
Allergies	Input
Asthma Medication Crisis	Input
Asthma Medication Daily	Input
BCG Vaccine	Input
BMI Class	Input
Bronchial Asthma	Input
Cough A Lot	Input
Daily Expectoration	Input
Dry Cough	Input
Ex Smoker	Input
Flu More Than Twice Year	Input
Flu Vaccine Usually	Input
Gender	Input
Height Class	Input
ID Test Table All	Key
Lacrimation Itch	Input
More Fatigue Than People Same Age	Input
Nose Sneeze	Input
Other Vaccines Prevention Respiratory Infections	Input
Pneumonia Vaccine	Input
Pneumonia	Input
Pulmonary Tuberculosis	Input
Rhinitis	Input
Shortness Breath	Input
Smoker	Input
String COPD	PredictOnly
Weight Class	Input
Wheezing Flu	Input
Wheezing Twelve Months	Input
Wheezing	Input

Figura 4.8 - Atributos utilizados na Árvore de Decisão

A Figura 4.9 ilustra a rede de dependências da AD. De entre os sintomas e os factores de risco colocados como input, o algoritmo conseguiu detectar a pieira nos últimos doze meses (*WheezingTwelveMonths*), a idade (*AgeClass*), a falta de ar (*ShortnessBreath*), a comichão e olhos lacrimejantes (LacrimationItch) e a asma brônquica (*BronchialAsthma*) como atributos com maior relação na previsão de DPOC.

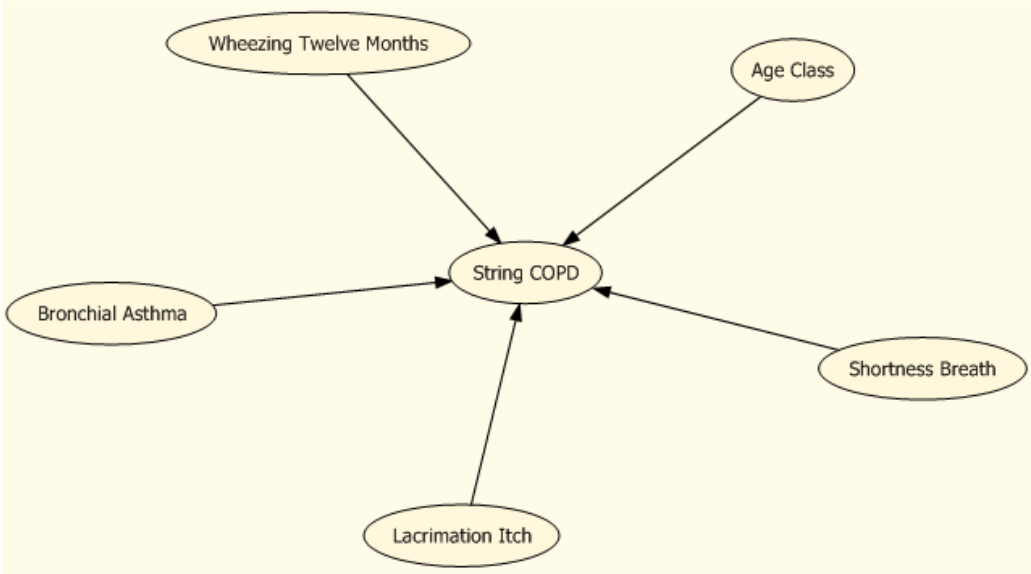


Figura 4.9 - *Decision Tree: Dependency Network*

De facto, como pode ser verificado na Figura 4.10 e na Tabela 4.1, verifica-se um aumento da percentagem de utentes que apresentam DPOC quando também têm falta de ar (*ShortnessBreath*) de 10.5% dos casos, para 28.7%. No ramo em que os utentes não têm falta de ar, a maior percentagem de incidência de DPOC é conseguida quando, em conjunção com este atributo, a idade do utente é superior a 65 anos e quando não sente comichão nem olhos lacrimejantes (26,3%). No ramo em que os utentes sentem falta de ar, a maior percentagem de incidência de DPOC é conseguida quando, em conjunção com este atributo, a idade do utente não está compreendida entre os 18 e os 40 anos, quando o utente teve pieira ou assobios nos últimos doze meses e quando tem asma brônquica (47,2%).

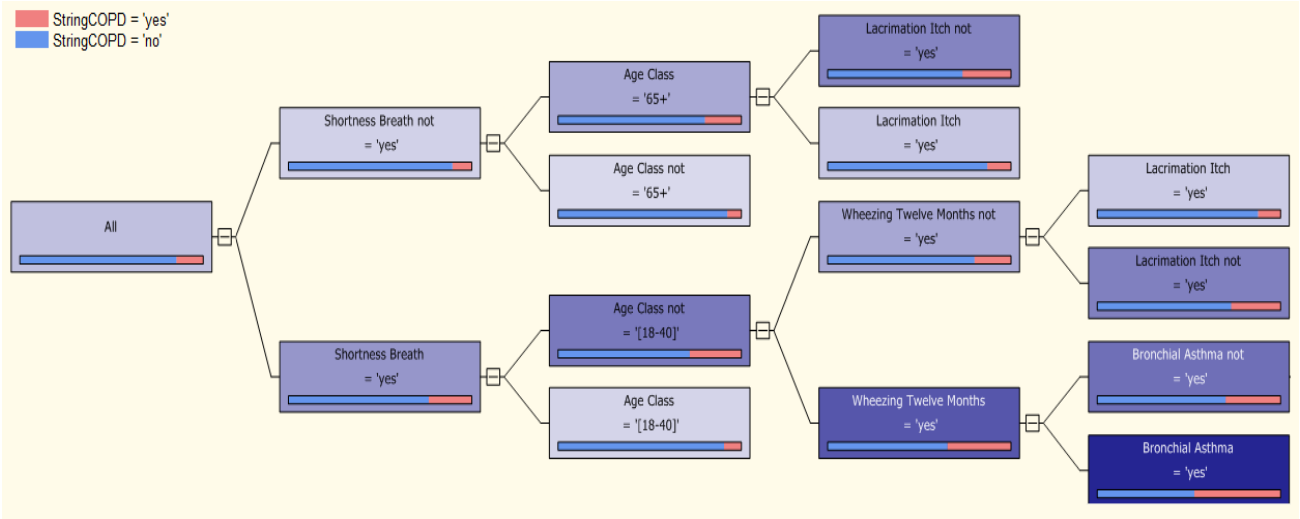


Figura 4.10 - Árvore de decisão para a DPOC

Condição	Casos Totais	Têm DPOC (Yes)		Não têm DPOC (No)	
		nº	%	nº	%
Shortness Breath not = 'yes'	826	87	10,5%	739	89,5%
ShortnessBreath not = 'yes' **Age Class not = '65+'**	643	51	7,9%	592	92,1%
Shortness Breath not = 'yes' **Age Class = '65+'**	183	36	19,7%	147	80,3%
Shortness Breath not = 'yes' **Age Class = '65+'** **Lacrimation Itch not = 'yes'**	95	25	26,3%	70	73,7%
Shortness Breath not = 'yes' **Age Class = '65+'** **Lacrimation Itch = 'yes'**	88	11	12,5%	77	87,5%
Shortness Breath = 'yes'	454	107	23,6%	347	76,4%
Shortness Breath = 'yes' **Age Class not = '[18-40]'**	335	96	28,7%	239	71,3%
Shortness Breath = 'yes' **Age Class = '[18-40]'**	119	11	9,2%	108	90,8%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months not = 'yes'**	140	28	20,0%	112	80,0%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months not = 'yes'** **Lacrimation Itch = 'yes'**	62	7	11,3%	55	88,7%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months not = 'yes'** **Lacrimation Itch not = 'yes'**	78	21	26,9%	57	73,1%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months = 'yes'**	195	68	34,9%	127	65,1%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months = 'yes'** **Bronchial Asthma not = 'yes'**	142	43	30,3%	99	69,7%
Shortness Breath = 'yes' **Age Class not = '[18-40]'** **Wheezing Twelve Months = 'yes'** **Bronchial Asthma = 'yes'**	53	25	47,2%	28	52,8%

Tabela 4.1 - *Decision Tree*: Resultados

O próximo modelo foi identificado com o algoritmo *Microsoft Naïve Bayes*. Este algoritmo verifica qual a influência que cada atributo tem individualmente na variável de previsão. Foram utilizados os mesmos parâmetros de entrada do que na AD. Desta vez, a rede de dependências mostra-nos oito atributos que têm maior relação na previsão de quem apresenta DPOC ou não: *ShortnessBreath*, *BronchialAsthma* e *WheezingTwelveMonths* (como na AD), *Wheezing* (se já alguma vez teve pieira ou assobios), *WheezingFlu* (pieira ou assobios com gripe), *MoreFatigueThanPeopleSameAge* (sente maior cansaço em relação às pessoas da mesma idade), *AsthmaMedicationDaily* (medicação para a asma diariamente) e *AsthmaMedicationCrisis* (medicação para a asma apenas em crises). Esta descrição pode ser visualizada na Figura 4.11.

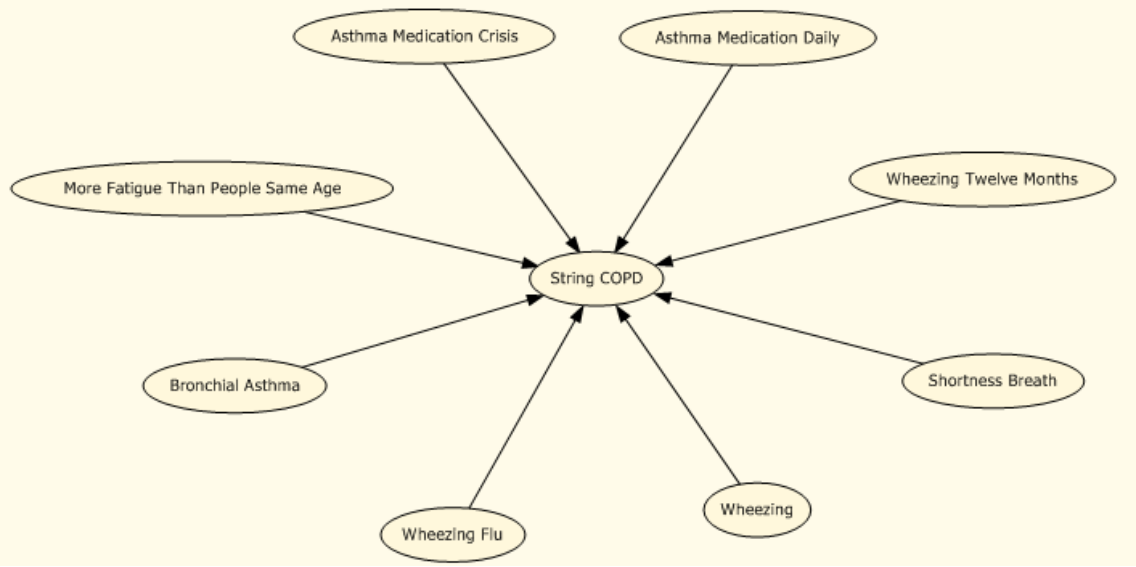


Figura 4.11 - *Naïve Bayes: Dependency Network*

A próxima análise com o algoritmo *Microsoft Naïve Bayes* mostra a distribuição dos valores de cada atributo pela variável de previsão (Figura 4.12).

Apesar de se verificar sempre (em todos os atributos) um aumento de incidências do sintoma quando o indivíduo apresenta DPOC, é de notar que, de entre os oito atributos que mais se relacionam com a previsão da DPOC, apenas o *MoreFatigueThanPeopleSameAge*, o *ShortnessBreath*, o *Wheezing* e o *WheezingTwelveMonths* é que conseguem apresentar uma maior incidência de casos positivos quando o indivíduo tem DPOC, com uma percentagem de, respectivamente, 56.7 %, 55.2 %, 67.0% e 55.2 %.

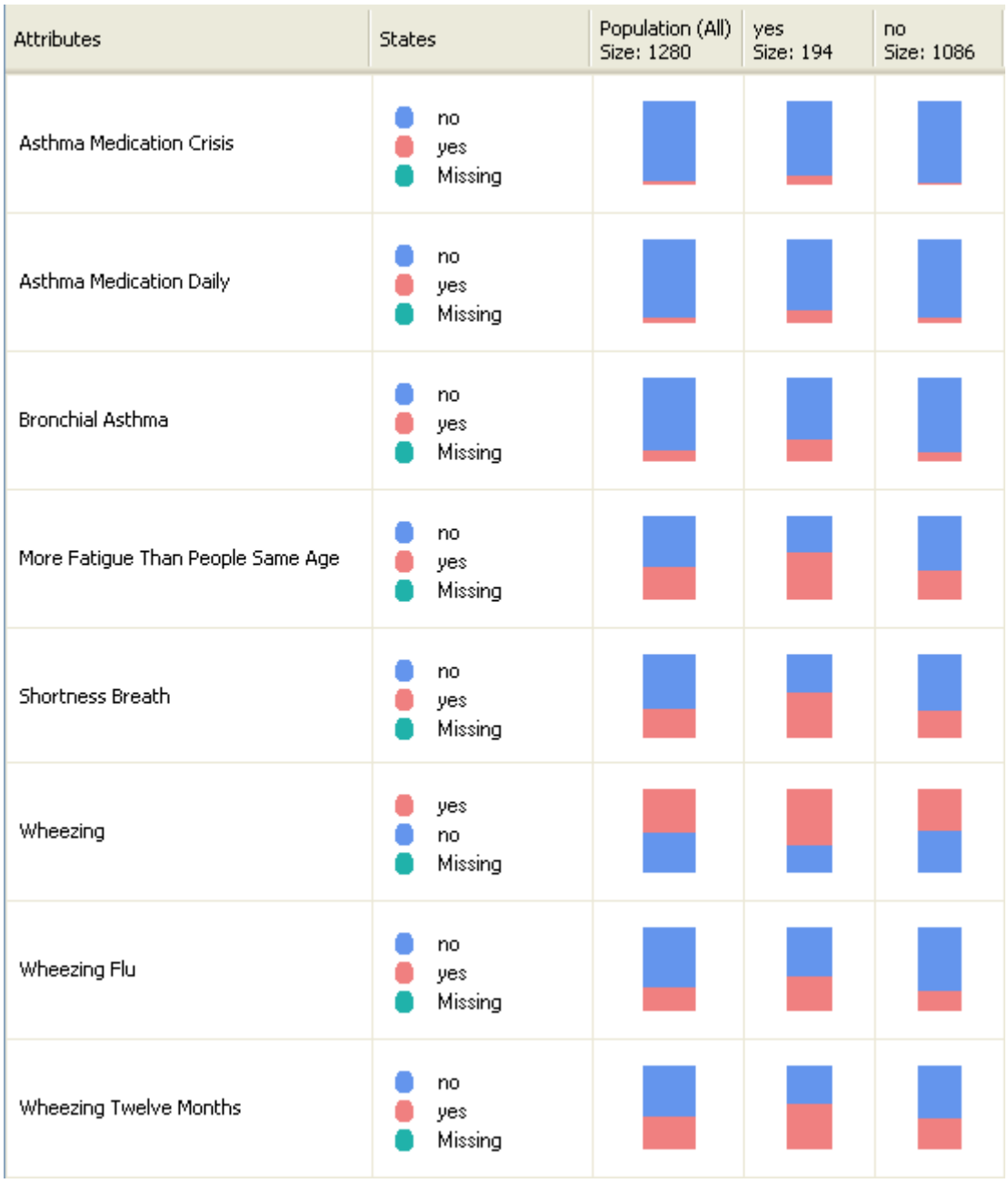
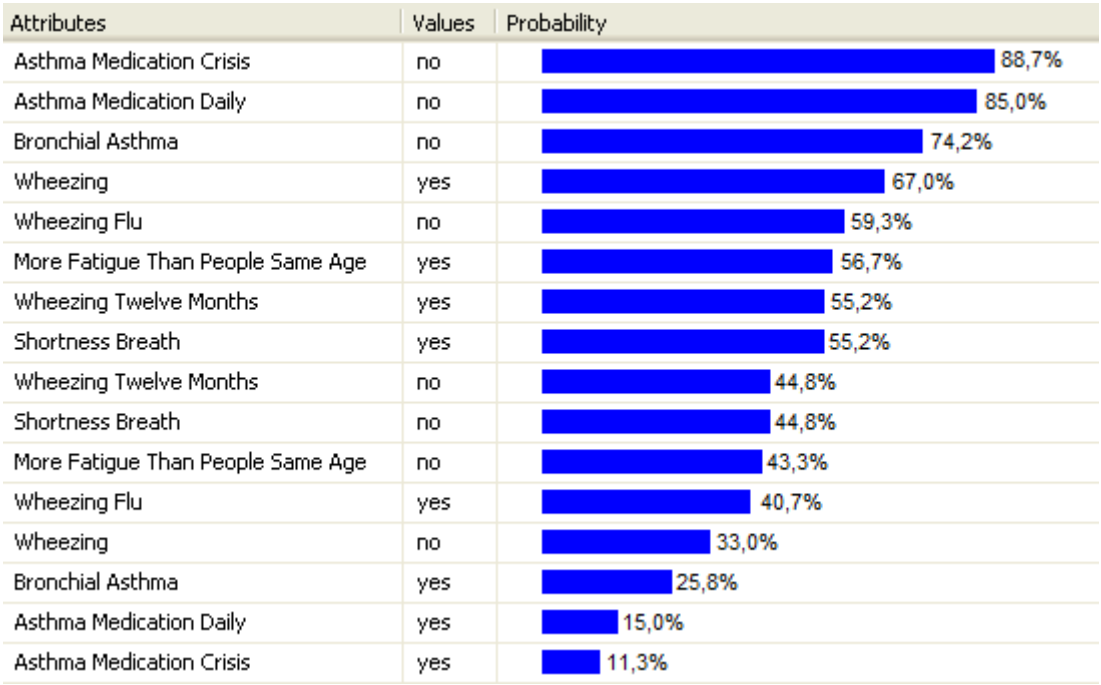


Figura 4.12 - *Naïve Bayes: Atribute Profiles*

Na Figura 4.13 podem ser vistas as características de cada atributo quando a variável de previsão *stringCOPD* (tem DPOC ou não) está com o valor "yes", ou seja, a probabilidade de um determinado valor do atributo estar presente em conjunto com a variável de previsão. Pode-se por exemplo concluir, com os dados disponíveis, que os utentes com DPOC, em 25,8% dos casos têm asma brônquica, em 55,2% têm falta de ar e em 67% têm pieira ou assobios.

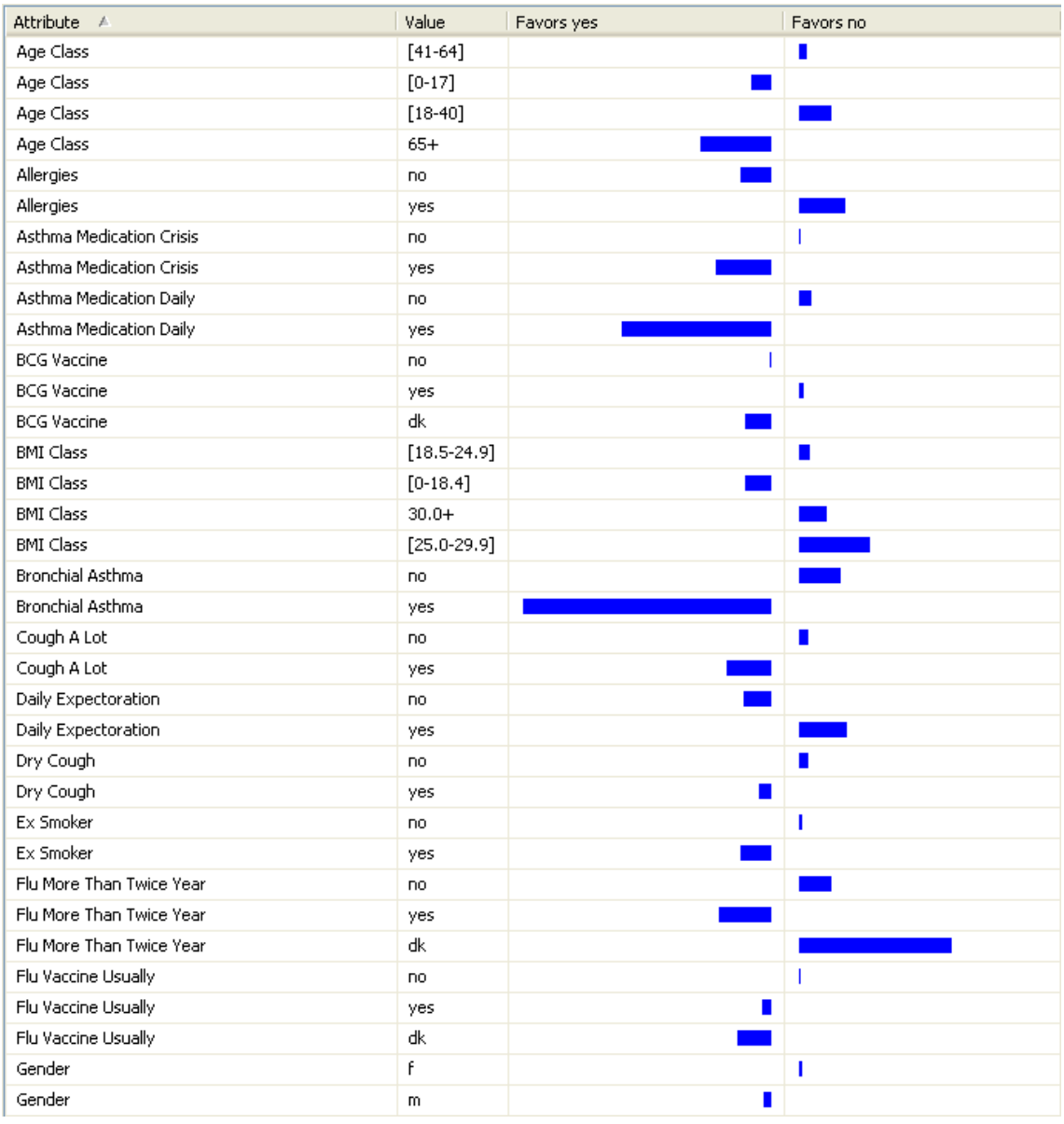


Attributes	Values	Probability
Asthma Medication Crisis	no	88,7%
Asthma Medication Daily	no	85,0%
Bronchial Asthma	no	74,2%
Wheezing	yes	67,0%
Wheezing Flu	no	59,3%
More Fatigue Than People Same Age	yes	56,7%
Wheezing Twelve Months	yes	55,2%
Shortness Breath	yes	55,2%
Wheezing Twelve Months	no	44,8%
Shortness Breath	no	44,8%
More Fatigue Than People Same Age	no	43,3%
Wheezing Flu	yes	40,7%
Wheezing	no	33,0%
Bronchial Asthma	yes	25,8%
Asthma Medication Daily	yes	15,0%
Asthma Medication Crisis	yes	11,3%

Figura 4.13 - *Naïve Bayes: Atribute Characteristics*

A próxima técnica utilizada para este primeiro objectivo de DM foi o *Microsoft Neural Network*. Os parâmetros de entrada utilizados foram os mesmos que foram utilizados anteriormente para as técnicas *Microsoft Decision Trees* e *Microsoft Naïve Bayes*. Esta técnica de DM utiliza um *Discrimination Viewer* que permite determinar as características que melhor prevêem os valores da variável de previsão (Figura 4.14 e Figura 4.15). Com isto, pode ver-se o que é que diferencia os utentes que sofrem de DPOC ("Favors yes") e os que não apresentam esta doença ("Favors no"). As características são ordenadas segundo uma pontuação que é calculada com base nos pesos de cada saída e a sua respectiva probabilidade de acontecimento. Por exemplo, para o atributo que nos diz se o utente sofre de asma brônquica (*BronchialAsthma*), o valor "yes" obteve uma pontuação de 100 (máxima), com um peso para o valor "yes" de 4.9 e para o valor "no" 0.49, e com uma probabilidade de acontecimento de 57.02% para o valor "yes" e de 42.87% para o valor "no". O atributo *AsthmaMedicationDaily*, que nos diz se o utente necessita de medicação diária para a asma, o valor "yes" obteve uma pontuação de 60.85, com um peso para o valor "yes" de 3 e para o valor "no" 0.74, apesar de ter uma probabilidade de acontecimento de 34.95% para o valor "yes" e de 64.94% para o valor "no".

É de notar que tanto a tuberculose pulmonar (*PulmonaryTuberculosis*) como a pneumonia (*Pneumonia*), quando apresentam o valor "yes", têm um peso maior nos utentes que sofrem de DPOC. Outra ocorrência interessante de apontar é o facto de o atributo *Allergies* (alergias) ter um peso maior nos utentes com DPOC quando apresenta o valor "no", e do atributo *Rhinitis* (rinite) ter um peso maior nos utentes com DPOC quando apresenta o valor "yes", apesar do peso do valor "dk" (*do not know* – não sabe) também apresentar um *score* elevado para os utentes que não têm DPOC. Para além disto, é de realçar também o facto de a classe de idades "65+" apresentar a maior pontuação das quatro classes de idades (quer para os "Favors yes" como para os "Favors no") e de ter um maior peso para o valor "yes" da variável de previsão.



Attribute	Value	Favors yes	Favors no
Age Class	[41-64]		
Age Class	[0-17]		
Age Class	[18-40]		
Age Class	65+		
Allergies	no		
Allergies	yes		
Asthma Medication Crisis	no		
Asthma Medication Crisis	yes		
Asthma Medication Daily	no		
Asthma Medication Daily	yes		
BCG Vaccine	no		
BCG Vaccine	yes		
BCG Vaccine	dk		
BMI Class	[18.5-24.9]		
BMI Class	[0-18.4]		
BMI Class	30.0+		
BMI Class	[25.0-29.9]		
Bronchial Asthma	no		
Bronchial Asthma	yes		
Cough A Lot	no		
Cough A Lot	yes		
Daily Expectoration	no		
Daily Expectoration	yes		
Dry Cough	no		
Dry Cough	yes		
Ex Smoker	no		
Ex Smoker	yes		
Flu More Than Twice Year	no		
Flu More Than Twice Year	yes		
Flu More Than Twice Year	dk		
Flu Vaccine Usually	no		
Flu Vaccine Usually	yes		
Flu Vaccine Usually	dk		
Gender	f		
Gender	m		

Figura 4.14 - *Neural Network: Discrimination Viewer* (1)

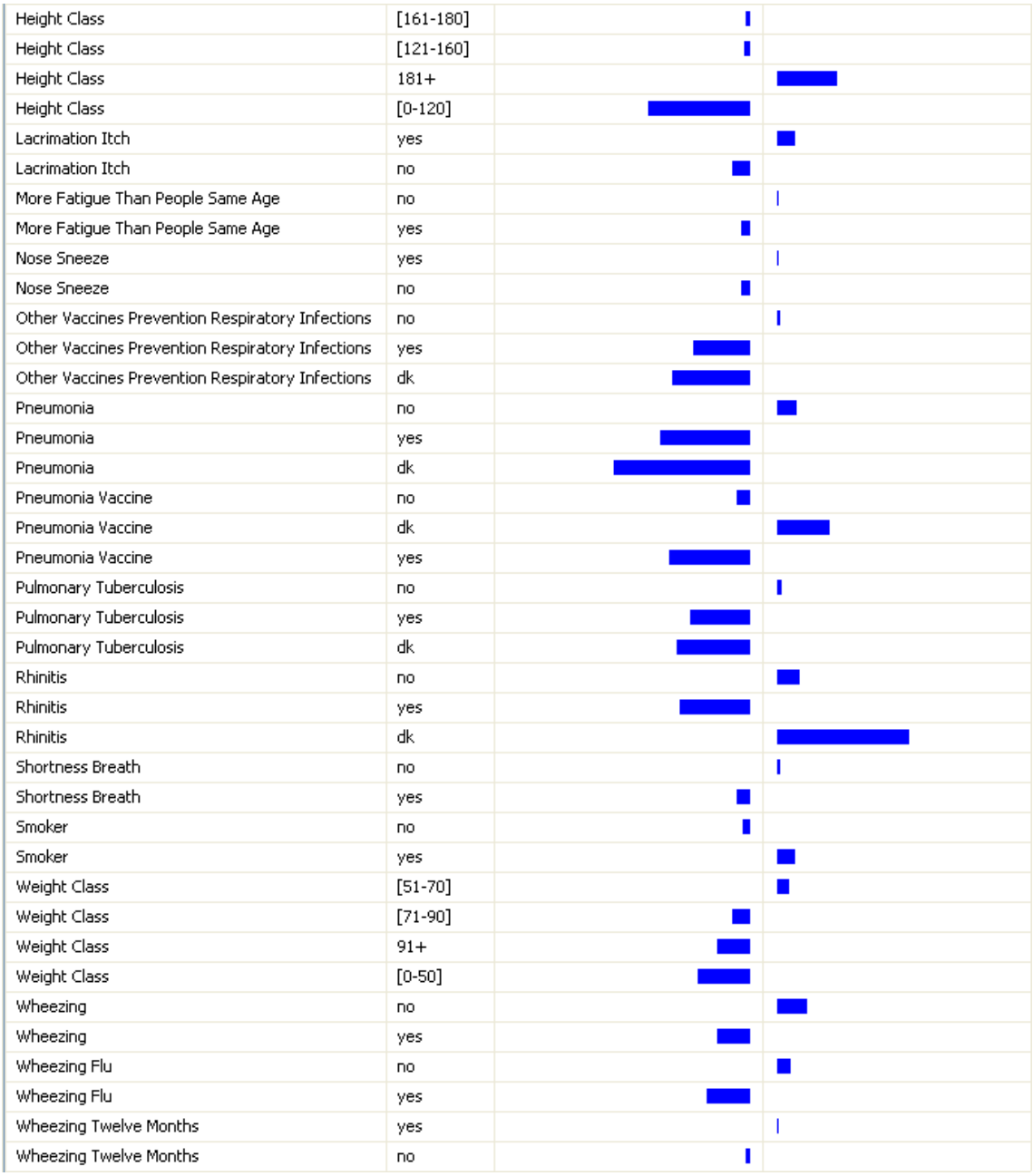


Figura 4.15 - *Neural Network: Discrimination Viewer* (2)

Depois de criados três modelos para este primeiro objectivo de DM, o próximo passo foi o de assegurar que os modelos faziam previsões com uma precisão considerada adequada. Para tal, foi utilizado o *Mining Accuracy Chart* do *Business Intelligence Development Studio* que fornece ferramentas como o *Lift Chart* e a *Classification Matrix* para determinar a precisão dos modelos de DM. O *Lift Chart* é um gráfico de linhas que mostra a performance dos modelos utilizando um conjunto de dados de teste. A *Classification Matrix* mostra o número correcto de previsões e o número de previsões erradas.

A Figura 4.16 ilustra o *Lift Chart* dos modelos construídos. Retrata quão bem (ou não) cada modelo prevê o número correcto de casos de DPOC no conjunto de dados de teste. A legenda do gráfico mostra as estatísticas de cada modelo de DM de uma determinada percentagem da população. Na Figura 4.16, a legenda contém informações para 50% da população de teste. Quando 50% da população é processada, a AD ("DT") conseguiu prever 45.07% dos casos correctamente, apresentando uma probabilidade de previsão de 90.08%. Para os mesmos 50% da população processada, a RNA ("NN1"), conseguiu prever 44.16% dos casos correctamente mas apresenta uma probabilidade de previsão de 88.04%, enquanto que a Naïve Bayes ("NB1") conseguiu prever correctamente 43.07% dos casos e tem uma probabilidade de previsão de 93.62%. Contudo, à medida que o processamento dos casos presentes do conjunto de testes continua, torna-se cada vez mais claro que a AD é o melhor modelo, seguido da RNA e finalmente da Naïve Bayes.

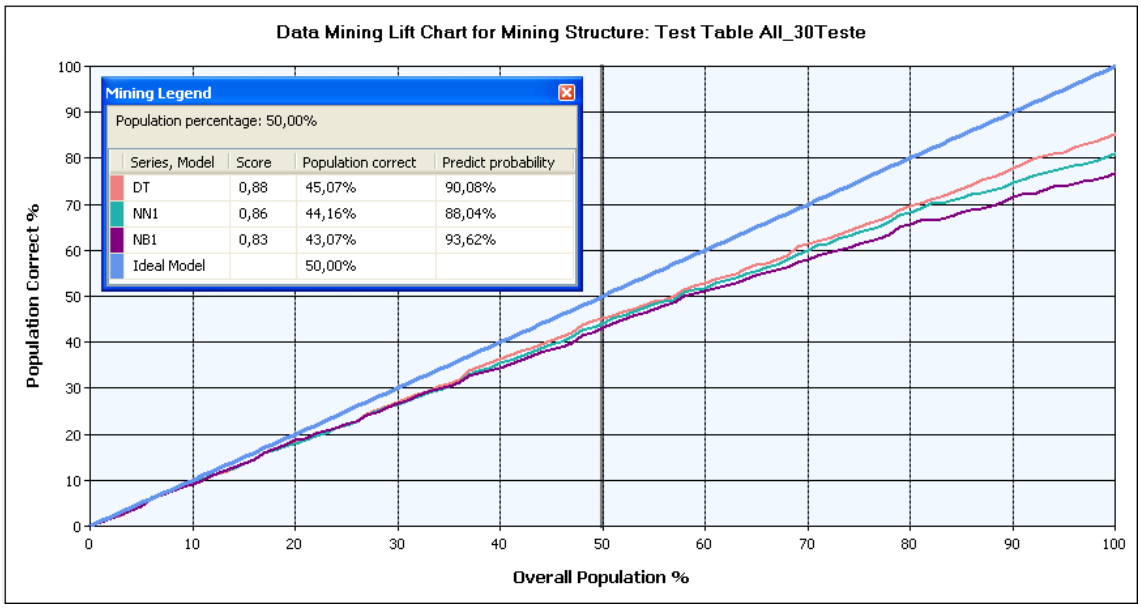


Figura 4.16 - *Lift Chart: Decision Tree*, *Naïve Bayes* e *Neural Network*

Sabe-se que os modelos de DM não são perfeitos nas suas previsões e que vão errar. A *Classification Matrix* (ou matriz de confusão) permite ver exactamente quais foram os erros que os modelos cometeram. A coluna da esquerda de cada grelha mostra o valor previsto pelo modelo. Na Figura 4.17 é apresentada a previsão dos casos que têm ("yes") ou não têm ("no") DPOC. As outras colunas mostram o valor real da variável.

Olhando para o primeiro modelo, por exemplo, pode-se dizer que:

- A primeira célula, que contém o valor 0, indica o número de verdadeiros positivos para o valor "yes". Este valor indica que o modelo conseguiu prever o valor correcto para os utentes que têm DPOC em 0 casos, isto porque o "yes" indica os utentes que têm essa doença;
- A célula que contém o valor 81 indica o número de falsos positivos, ou seja, quantas vezes é que o modelo previu que o utente não tinha DPOC quando na realidade apresentava esta doença;
- A célula da direita, na primeira linha, que contém o valor 0 indica o número de falsos positivos para o valor "no". Este valor indica que o modelo previu que o utente tinha DPOC quando na realidade não tem, 0 vezes, isto porque o "no" indica os utentes que não têm a doença;
- Finalmente, a célula que contém o valor 467 indica o número de verdadeiros positivos para o valor "no" da variável de previsão. Por outras palavras, em 467 casos, o modelo previu correctamente que o utente não tinha DPOC.

É possível determinar a precisão total do modelo, dividindo a soma dos valores da diagonal principal pelo somatório dos valores de todas as células. Então, o primeiro modelo (AD) apresenta uma precisão de $467 \div (81 + 467) = 85{,}22\%$. O segundo modelo (RN) tem uma precisão de $(429 + 15) \div (15 + 38 + 66 + 429) = 81{,}02\%$. O terceiro modelo (Naïve Bayes) apresenta uma precisão global de $(25 + 395) \div (25 + 72 + 56 + 395) = 76{,}64\%$.

Counts for DT on String COPD:

Predicted	yes (Actual)	no (Actual)
yes	0	0
no	81	467

Counts for NN1 on String COPD:

Predicted	yes (Actual)	no (Actual)
yes	15	38
no	66	429

Counts for NB1 on String COPD:

Predicted	yes (Actual)	no (Actual)
yes	25	72
no	56	395

Figura 4.17 - *Classification Matrix*: *Decision Tree*, *Naïve Bayes* e *Neural Network*

No entanto, pode-se concluir que todos os modelos são mais eficientes em prever os casos em que o utente não apresenta DPOC do que em prever os casos em que o utente sofre efectivamente de DPOC. Isto acontece muito provavelmente porque a amostra não está balanceada (muito mais dados de utentes que não têm DPOC do que de utentes que apresentam esta doença), influenciando assim os resultados. Algumas ferramentas de DM fornecem estratégias/métodos para balancear os dados. Todavia, a ferramenta utilizada, o *SQL Server Business Intelligence Development Studio*, não permite este tipo de manipulação do conjunto de dados, nem escolher qual o conjunto de dados que se quer para treino e para teste, possibilitando apenas a divisão aleatória do conjunto como um todo, mediante a percentagem requerida.

Para o próximo problema estudado, a segmentação dos utentes que apresentam DPOC, foi utilizado o *Microsoft Clustering* como técnica de DM. O conjunto de dados utilizado nesta análise difere do conjunto de dados usado no problema de classificação. O que se pretende agora é descobrir padrões nos dados, com uma tarefa não supervisionada, e não fazer qualquer tipo de previsão. Portanto, o conjunto de dados utilizado contém unicamente os dados dos utentes que apresentam DPOC, excluindo os dados dos utentes que não têm esta doença. Com esta partição dos dados, o conjunto de dados para a análise deste problema apresenta um total de 275 registos. A Figura 4.18 mostra os atributos de *input* que foram considerados.

Para analisar o problema em causa, foi necessário proceder a alterações nos parâmetros do algoritmo. Mais concretamente, o valor da variável "CLUSTER_COUNT" do *Microsoft Clustering,* que por defeito é "10", foi alterado para "0". Esta variável especifica o número aproximado de *clusters* que são construídos pelo algoritmo. Se o número aproximado de *clusters* não pode ser construído a partir do conjunto de dados, o algoritmo constrói tantos *clusters* quanto possível. Definindo o parâmetro "CLUSTER_COUNT" a 0 faz com que o algoritmo use um procedimento heurístico para determinar da melhor maneira o número de *clusters* para construir.

Structure	Cluster
	Microsoft_Clustering
Age Class	Input
Allergies	Input
Asthma Medication Crisis	Input
Asthma Medication Daily	Input
BCG Vaccine	Input
BMI Class	Input
Bronchial Asthma	Input
Cough A Lot	Input
Daily Expectoration	Input
Dry Cough	Input
Ex Smoker	Input
Flu More Than Twice Year	Input
Flu Vaccine Usually	Input
Gender	Input
Height Class	Input
ID Test Table	Key
Lacrimation Itch	Input
More Fatigue Than People Same Age	Input
Nose Sneeze	Input
Other Vaccines Prevention Respiratory Infections	Input
Pneumonia Vaccine	Input
Pneumonia	Input
Pulmonary Tuberculosis	Input
Rhinitis	Input
Shortness Breath	Input
Smoker	Input
Weight Class	Input
Wheezing Flu	Input
Wheezing Twelve Months	Input
Wheezing	Input

Figura 4.18 - Atributos utilizados no *Clustering*

Foram obtidos quatro *clusters* no total. Analisando a primeira parte dos atributos (Figura 4.19), no *cluster* com maior população, é de realçar que mais de metade (53.6%) da população pertencente a este grupo tem mais de 65 anos. Também é no *cluster* 1 que existe maior incidência de utentes com alergias (60.6%) e de utentes que necessitam de medicação para a asma (20.7% só em crise e 33.2% diariamente), sendo também, naturalmente, o *cluster* que apresenta maior incidência de utentes com asma brônquica (53.7%). O *cluster* 1 também é o grupo de utentes que tem maior incidência de utentes com muita tosse (47.5%) e com expectoração diária (50.1%).

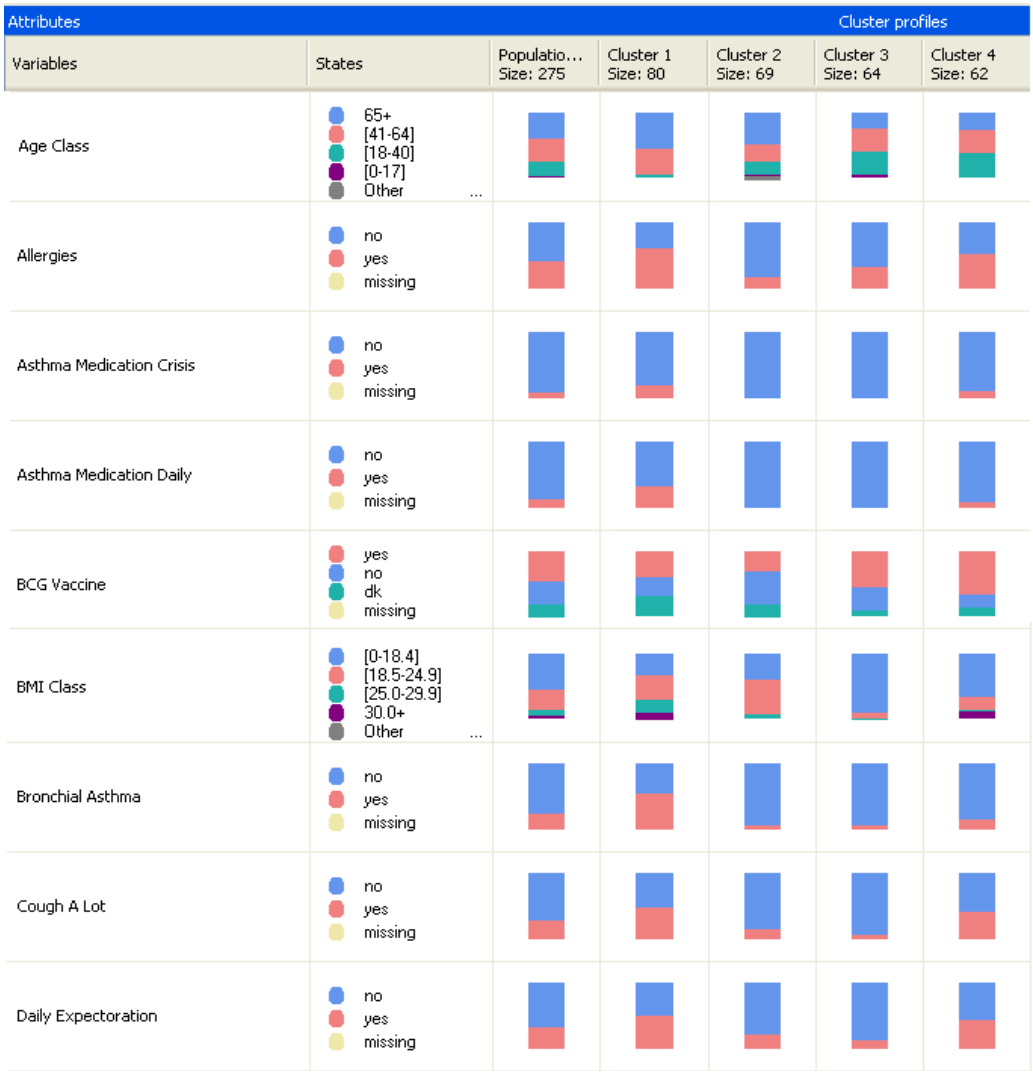


Figura 4.19 - *Clustering: Cluster Profiles* (1)

Na segunda parte dos atributos presentes nos diferentes *clusters* (Figura 4.20), é de notar que é também no *cluster* 1 que existe uma maior incidência de utentes que são ex-fumadores (37.9%). De realçar que é no *cluster* 4 que existe a maior incidência de utentes com tosse seca (86.7%). Todavia, o *cluster* 1 não fica muito atrás apresentando uma incidência deste atributo de 77.6%. Neste agrupamento de utentes mais populado, também se pode verificar que este apresenta uma maior incidência de utentes que têm gripe mais de duas vezes por ano (57.9%), costumam vacinar-se contra a gripe (62.3%), têm comichão e olhos lacrimejantes (68.7%), sentem maior cansaço em relação a pessoas da mesma idade (82.7%) e têm corrimento nasal e espirros quando não estão constipados nem com gripe (91.0%).

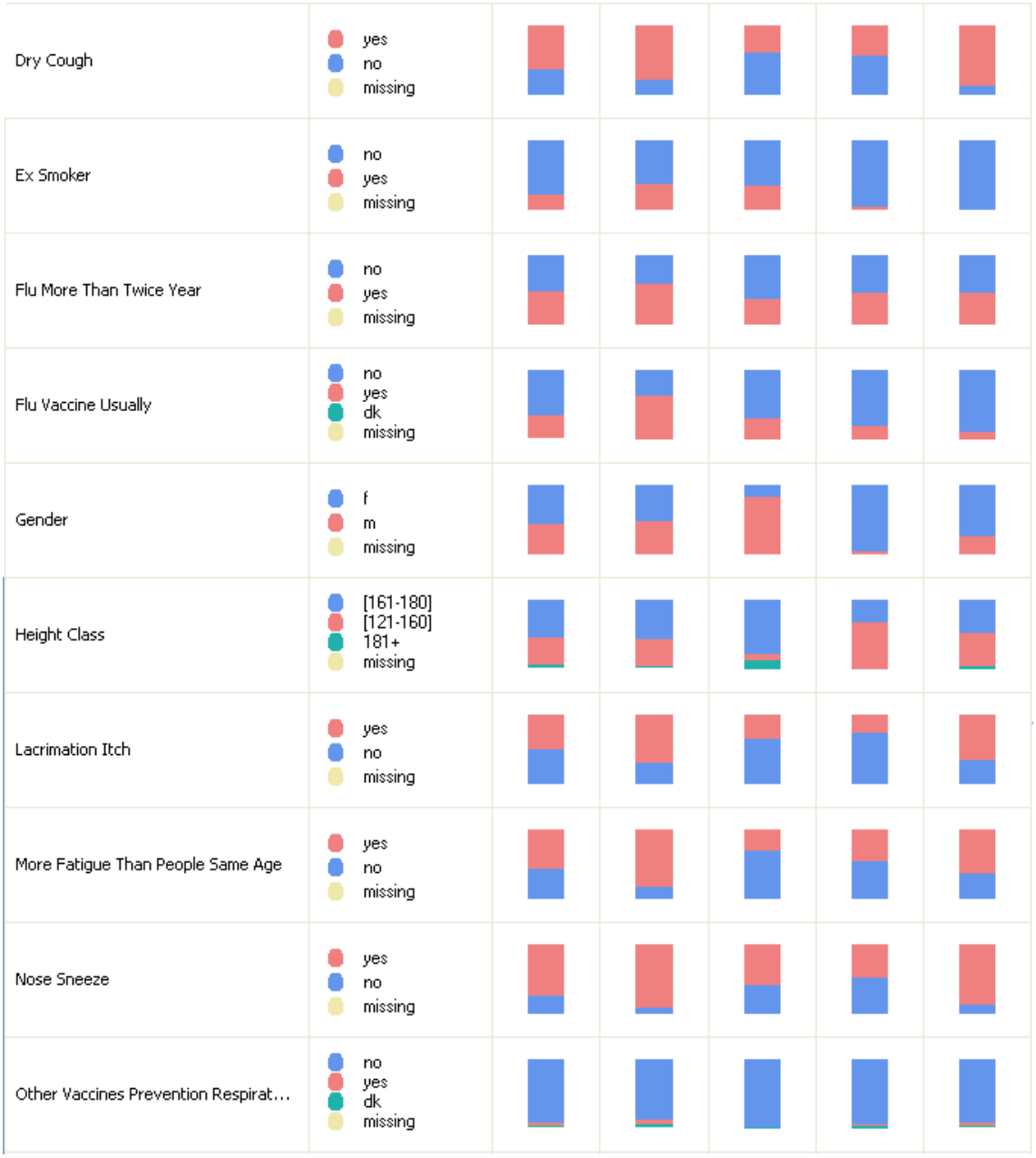


Figura 4.20 - *Clustering: Cluster Profiles* (2)

Analisando a última parte dos atributos presentes nos diferentes *clusters* (Figura 4.21), realça-se o facto de que o *cluster* 1 continua a apresentar a maior incidência de casos positivos da maioria dos atributos presentes na análise. De notar por exemplo a maior incidência de utentes com pneumonia (33.2%), tuberculose pulmonar (12.0%), rinite (43.1.%), falta de ar (81.9%). Contrastando com estas variáveis, está o atributo que indica se o utente é fumador, apresentando quase a totalidade (99.2%) dos utentes deste *cluster* como sendo não fumadores.

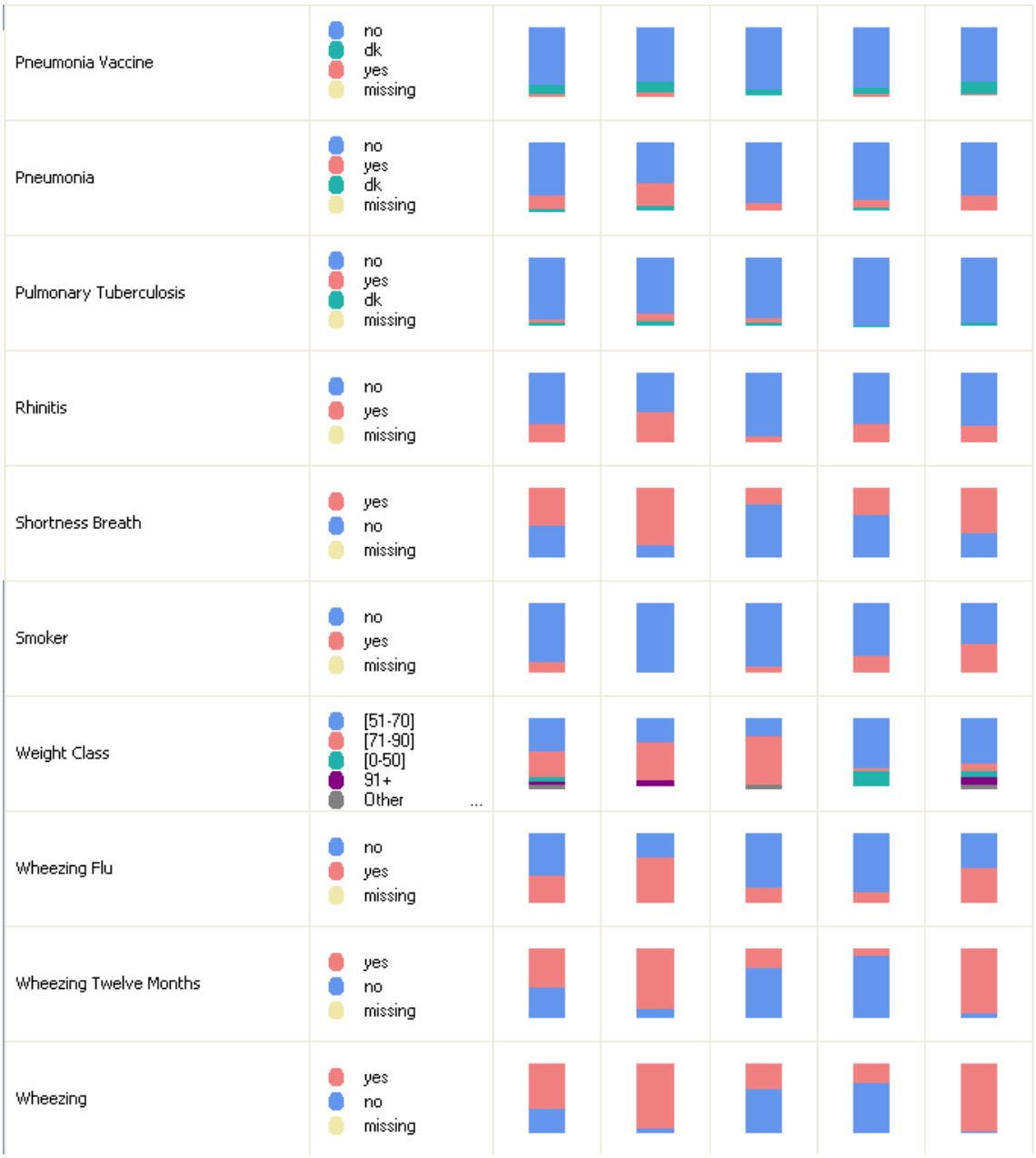


Figura 4.21 - *Clustering: Cluster Profiles* (3)

Figura 4.22 - *Clustering*: Diagrama dos *clusters* e das ligações *inter-clusters*

É de realçar ainda que o *cluster* 1 e o *cluster* 4 (representando em conjunto mais de metade dos utentes analisados) apresentam resultados muito próximos (ver Figura 4.22) diferenciando-se apenas nos seguintes aspectos:

- Na variável que indica a incidência da asma brônquica, o *cluster* 1 tem 53.7% de casos positivos enquanto que o *cluster* 4 tem apenas 16.7% de casos positivos (conseguindo, ainda assim, apresentar uma maior incidência do que o *cluster* 2 e o *cluster* 3);
- No sexo dos utentes, o *cluster* 1 tem uma distribuição mais equitativa enquanto que o *cluster* 4 apresenta uma incidência de 73.2% de utentes do sexo feminino;
- Nas duas variáveis que representam a relação dos utentes com o tabaco, os dois *clusters* apresentam resultados opostos. No *cluster* 1, os utentes que fumam representam 0.08% do grupo (a menor incidência deste atributo dos quatro *clusters* identificados), enquanto que no *cluster* 4, os utentes que fumam representam 40.9% do grupo (a maior incidência deste atributo dos quatro *clusters* identificados). Em relação aos ex-fumadores, o *cluster* 1 tem uma incidência de 37.9% de casos positivos (a maior incidência deste atributo dos quatro *clusters* identificados), enquanto que o *cluster* 4 tem apenas 2.1% de casos positivos (a menor incidência deste atributo dos quatro *clusters* identificados);
- No seguimento dos atributos que caracterizam a relação dos utentes com o tabaco, os resultados obtidos para a variável que indica se os utentes com DPOC costumam vacinar-se contra a gripe também apresenta resultados opostos nos dois *clusters*. Enquanto que no *cluster* 1 62.3% dos utentes se vacinam regularmente contra a gripe, no *cluster* 4, apenas 12.2% dos utentes o fazem.

Capítulo 5 – Conclusões

Neste capítulo final, para além de ser apresentada uma síntese do trabalho realizado, são também realçadas as principais contribuições do projecto. O capítulo termina com a apresentação de algumas propostas de trabalho futuro.

5.1 Síntese do Trabalho Realizado

Este trabalho apresentou um sistema de BI que permite realizar análises aos dados recolhidos pela FPP, identificando padrões de incidência e modelos preditivos da DPOC, auxiliando assim a tomada de decisão dos profissionais de saúde desta área.

Os objectivos delineados para este trabalho passaram por, primeiramente, conceber e implementar uma aplicação *Web* e sua respectiva BDO para a introdução dos dados dos utentes da FPP e para a gestão dos mesmos. Foi necessário definir a arquitectura do sistema de BI a implementar e caracterizar os seus principais componentes, assim como as tecnologias a utilizar. Outro dos objectivos traçados para este trabalho foi a definição da arquitectura (modelo de dados) do DW que serve de suporte ao armazenamento de dados e a sua respectiva implementação. Por fim, foram feitas análise aos dados presentes no DW recorrendo a tecnologias e técnicas avançadas como o OLAP e os algoritmos de DM para identificação de padrões e tendências nos dados.

O conjunto de dados disponibilizado pela FPP apresenta dados de 1898 utentes. Os dados presentes nesse conjunto de dados correspondem às respostas de cada utente a uma série de perguntas de um questionário e aos resultados de um exame médico, denominado espirometria.

A BDO que suporta os dados da aplicação *Web* contém doze tabelas, incluindo as que armazenam os dados pessoais do utente, as que permitem guardar os dados relacionados com o inquérito realizado e o exame espirométrico, e as que possibilitam a autenticação à aplicação. De realçar ainda que os campos mais críticos se encontram devidamente encriptados. É o caso do nome do utente, por razões éticas e de confidencialidade, e da *password* de autenticação à aplicação *Web*, por questões de segurança. A aplicação *Web* possibilita a autenticação de três tipos de utilizadores (com algumas tarefas distintas) e inclui um inquérito aos utentes melhorado e com mais questões do que aquele que foi apresentado com os dados iniciais disponibilizados, assim como a gestão por parte dos administradores das tabelas críticas do sistema.

O modelo de dados do DW é composto por um vector de análise (esquema em estrela), representado pela tabela de factos "FactFPP". Esta tabela está ligada a várias tabelas dimensões que permitem a análise desses dados sob diversas perspectivas: "*Time*" (Tempo), "*Location*" (Localização), "*Profession*" (Profissão), "*Patient*" (Utente), "*Smoke Characterization*" (Caracterização do Fumo), "*Allergy Characterization*" (Caracterização das Alergias), "*Cough Characterization*" (Caracterização da Tosse), "*Fatigue Characterization*" (Caracterização do Cansaço) e "*Pulmonary Diseases Characterization*" (Caracterização das Doenças Pulmonares). As dimensões das caracterizações são tabelas onde estão agrupadas todas as respostas associadas às perguntas que estão relacionadas com um determinado tema. De realçar o facto de que o modelo foi concebido de maneira a ser possível o acréscimo de mais vectores de análise.

Dos diversos problemas que foram identificados nos dados disponibilizados pela FPP, realçam-se os seguintes: vários caracteres trocados, respostas não padronizadas, registos em branco, respostas nos atributos errados, a necessidade de calcular valores e a necessidade de separar alguns atributos em dois atributos distintos. Foi por exemplo o caso do atributo "Tabaco", que foi separado em "Fumador" e "Ex-Fumador" e da "Tuberculose Pulmonar", onde se criou um atributo específico para a idade na qual os utentes tiveram esta doença. Depois da limpeza e da transformação do conjunto de dados inicial, o número total de registos passou a ser de 1828.

As várias análises realizadas aos dados reforçaram essencialmente a ideia de que a DPOC é uma doença difícil de diagnosticar sem o exame espirométrico. Isto acontece principalmente porque, apesar de se terem confirmados alguns factores de risco relacionados

com a DPOC e identificados alguns padrões desta doença, parece haver algum tipo de paradoxo com os seus sintomas. Estas situações aparecem por exemplo nas análises feitas recorrendo à componente OLAP com a caracterização das respostas obtidas ao grupo de questões relacionadas com o cansaço e com a tosse.

A variação dos algoritmos que foram utilizados nas análises com DM teve como base os dois principais objectivos de análise. Para um problema de classificação, prever quem tem DPOC atendendo aos sintomas e factores de risco, foram utilizadas as seguintes técnicas de DM: *Microsoft Decision Trees*, *Microsoft Naïve Bayes* e *Microsoft Neural Network*. Para um problema de segmentação, descobrir os padrões escondidos nos dados dos utentes que apresentam DPOC, foi usado o *Microsoft Clustering*.

Também se verificou que os modelos identificados para a previsão da DPOC, recorrendo aos algoritmos de DM, são mais eficientes em prever os casos em que o utente não apresenta a doença do que em prever os casos em que o utente sofre efectivamente da doença. Estes resultados podem ser derivados do facto da amostra não estar balanceada, apresentando muitos mais dados de utentes que não têm a doença do que de utentes que apresentam esta doença, influenciando assim os resultados.

5.2 Contribuições

Os objectivos delineados para este trabalho de dissertação tiveram como plano de fundo contribuir para a resolução de um problema real, a situação epidemiológica da DPOC em Portugal, construindo um sistema de BI para o estudo desta doença pulmonar.

A área da Saúde, assim como outras áreas de aplicação de BI, apresenta necessidades ao nível do armazenamento e análise de dados para processar a tomada de decisão. Por ser uma área crítica, necessita não só que a informação seja processada o mais rapidamente possível mas também e, fundamentalmente, com o maior grau de precisão possível.

Este trabalho mostrou ser um bom caso prático da utilidade destes sistemas, num ambiente e numa organização que apresentam problemas reais.

O trabalho desenvolvido mostrou ainda que, com os sistemas de BI, torna-se mais fácil extrair conhecimento através dos dados e informação que se encontram espalhados pela organização, transformando e carregando estas várias fontes de dados para um DW. O modelo de DW proposto neste trabalho permite a análise da DPOC, tendo como base os dados recolhidos pela FPP. Todavia, para além de permitir a análise desta doença, espera-se que o contributo do modelo de DW proposto seja ainda maior, podendo no futuro ser expandido para o estudo de outras doenças pulmonares, inseridas no âmbito da FPP.

A recolha de dados dos utentes numa simples folha de cálculo, muitas vezes com erros e com dados incompletos, não permite uma análise precisa e profunda da informação recolhida. Portanto, a concepção e implementação de uma aplicação *Web* e da sua respectiva BDO vem permitir que os dados dos utentes, das respostas aos inquéritos e dos exames espirométricos sejam inseridos respeitando uma série de normas e restrições, armazenando-os de uma forma padronizada. Um dos outros contributos deste trabalho é, naturalmente, a aplicação de técnicas OLAP e de algoritmos de DM numa área aplicacional que até agora estava condicionada a um conjunto mais restrito de análises que não permitia identificar padrões nos dados nem modelos preditivos da DPOC.

Como se trata de um estudo novo que não tem precedentes a nível nacional, o sistema de BI para a DPOC concebido e implementado neste trabalho poderá trazer informação útil e contribuir para a formulação inicial de novo conhecimento na área. Apenas com informação acerca dos padrões de incidência e das possíveis causas, mas também com a identificação de modelos preditivos, será possível combater a DPOC e aplicar melhores políticas de saúde, e como esta doença pulmonar é uma doença muito ligada a comportamentos, é possível intervir e prevenir.

É de realçar ainda que, no decurso deste trabalho, foi efectuada a seguinte publicação numa conferência internacional da área:

> Dinis, R., Ribeiro, A., Santos, M. Y., Cruz, J., Araújo, A. "*A Business Intelligence Infrastructure Supporting Respiratory Health Analysis*", Proceedings the First International Conference on Business Intelligence and Technology (BusTECH'2011), 25-30 September 2011, Rome, Italy (Published by XPS), ISBN: 978-1-61208-160-1, pp. 13-19 (URI: http://hdl.handle.net/1822/13953).

5.3 Propostas de Trabalho Futuro

Em relação a trabalhos futuros, será interessante a realização de análises com mais dados de utentes que apresentam DPOC. Só assim será possível não só analisar os dados com maior rigor, mas também identificar mais padrões de incidência e modelos preditivos para a DPOC do que os resultados apresentados neste trabalho. Havendo a necessidade de uma apropriada caracterização epidemiológica da situação em Portugal, espera-se ainda que os dados que vão ser futuramente recolhidos com a utilização da aplicação *Web* concebida e desenvolvida para o efeito englobem regiões de todo o país, continente e ilhas inclusive (onde, por exemplo, as condições climatéricas são diferentes). Assim, é possível melhor caracterizar geograficamente a doença, uma vez que o conjunto de dados disponibilizado apresentava, na sua maioria, utentes da zona da Grande Lisboa, mas também do Grande Porto (isto pode ser explicado por serem as duas regiões mais populadas em Portugal, mas também denota um conjunto de dados não balanceado com incidências muito elevadas na região da Grande Lisboa).

Também está prevista a criação de um processo de ETL automatizado entre a BDO da aplicação *Web* e o DW. Para as futuras análises dos dados que serão armazenados na BDO da aplicação *Web*, será vantajoso utilizar a ferramenta *Integration Services* da *Microsoft.* De facto, tal permitirá criar processos automatizados e controlar o fluxo das tarefas necessárias para todos os passos de ETL requeridos. Este processo permitirá a extracção, limpeza, transformação e carregamento dos dados da BDO para o DW, mas principalmente o seu refrescamento, sem qualquer esforço adicional.

O modelo do DW apresentado foi desenhado a pensar na evolução do esquema em estrela para um esquema em constelação, à medida que novas tabelas de factos vão sendo adicionadas e o modelo ser expandido para o estudo de outras doenças pulmonares inseridas no âmbito da FPP. Neste momento, está previsto o acréscimo de duas tabelas de factos: uma para o estudo da pneumonia e outra para o estudo do cancro do pulmão. Com o crescimento da constelação, mais sintomas e dados sobre os indivíduos podem ser relacionados no estudo de uma ou mais doenças.

Por último, os outros trabalhos futuros cogitados foram a definição de KPI para a DPOC e a implementação de *dashboards* com a inclusão dos indicadores de desempenho definidos. A inclusão de *dashboards* no sistema de BI com as medições dos KPI vai ter como finalidade principal servir de plataforma de visualização de suporte analítico directo para as tomadas de decisão da FPP, monitorizando os principais indicadores de uma maneira simples e elegante visualmente e permitindo a percepção quase de forma imediata de uma determinada situação.

Referências

Alter, S. (1999). *Information Systems: A Management Perspective*: Addison Wesley Longman.

Ariyachandra, T., & Watson, H. J. (2005). Key Factors in Selecting a Data Warehouse Architecture. *Business Intelligence Journal, 10*(2).

Ariyachandra, T., & Watson, H. J. (2006). Wich Data Warehouse Architecture Is Most Successful? *Business Intelligence Journal, 11*(1).

Berry, M., & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. USA: John Wiley and Sons, Inc.

Berry, M., & Linoff, G. (2004). *Data Mining Techniques for Marketing, Sales and Customer Relationship Management* (2 ed.). USA: Wiley Publishing, Inc.

Berson, A., Smith, S., & Thearling, K. (2000). *Building Data Mining Applcations for CRM*. USA: McGraw-Hill.

Chang, G., Healey, M., McHugh, J., & Wang, J. (2001). *Mining the World Wide Web: An Information Search Approach*. USA: Kluwer Academic Publishers.

Chapman, P., Clinton, J., Kerber, R., Khabanza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 - Step-by-step Data Mining Guide. *CRISP-DM Consortium*.

Chaudhuri, S., & Dayal, U. (1997). An overview of Data Warehousing and OLAP technology. *SIGMOD Rec, 26*(1), 65-74.

Cody, W. F., Kreulen, J. T., Krishna, V., & Spangler, W. S. (2002). The integration of business intelligence and knowledge management. *IBM Systems Journal 41*, 697-713.

Cortez, P. (2002). *Modelos inspirados na Natureza para a Previsão de Séries Temporais.* Tese de Doutoramento, Departamento de Informática, Universidade do Minho.

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Eckerson, W. (2003a). Four Ways to Build a Data Warehouse. *What Works: Best Practices in Business Intelligence and Data Warehousing, 15*.

Eckerson, W. (2003b). Smart Companies in the 21st Century: The Secret of creating Successful Business Intelligent Solutions. *The Data Warehouse Institute*.

Eckerson, W. (2005). Data Warehouse Builders Advocate for Diferent Architectures. *Application Development Trends*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37-54.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurasamy, R. (1996). Advances in Knowledge Discovery and Data Mining. *AAAI Press / MIT Press*.

Gardner, S. R. (1998). Building the Data Warehouse. *Communications of the ACM, 41*(9), 52-60.

GOLD. (2010). *Global Strategy for the Diagnosis, Management and prevention of Chronic Obstructive Pulmonary Disease*. Global Initiative for Chronic Obstructive Lung Disease, Medical Communications Retrieved from http://www.goldcopd.com.

Golfarelli, M., & Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies*. San Francisco: McGraw-Hill Osborne Media.

Groth, R. (2000). *Data Mining: Building Competitive Advantage*. USA: Prentice Hall.

Hagan, M. T., Demuth, H. B., & Beale, M. (1996). *Neuronal Network Design*: PWS Publishing Company.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*: Morgan Kaufmann Publishers.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2 ed.): Morgan Kaufmann Publishers.

Inmon, W. H. (2005). *Building the Data Warehouse*. New York: Wiley.

Kartalopoulos, S. (1996). Understanding Neural Network and Fuzzy Logic - Basic Concepts and Applications. *IEEE Press*.

Kimball, R., Reeves, L., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing, and Deploying Data Warehouses*: John Wiley & Sons.

Loveridge, B., West, P., Kryger, M. H., & Anthonisen, N. R. (1986). Alteration in breathing pattern with progression of chronic obstructive pulmonary disease. *American Review of Respiratory Disease, 134*(5).

McCarthy, K., & Dweik, R. A. (2010). *Pulmonary Function Testing*. eMedicine Pulmonology Retrieved from http://emedicine.medscape.com/article/303239-overview.

Michalsky, R. S., Bratko, I., & Miroslav, K. (1998). *Machine Learning and Data Mining Methods and Applications*. England: John Wiley and Sons, Inc.

Moody, D. L., & Kortink, M. (2003). From ER Models to Dimensional Models: Bridging the Gap between OLTP and OLAP Design, Part I. *Business Intelligence Journal, 8*(3), 7-24.

ONDR. (2009). Relatório do Observatório Nacional das Doenças Respiratórias 2009 from http://www.ondr.org/relatorios_ondr.html

ONDR. (2011). Dia mundial da D.P.O.C. , from http://www.ondr.org/dpoc.htm

Pareek, D. (2006). *Business Intelligence for telecommunications*. New York: Auerbach Publications.

Quinlan, J. R. (1998). *C4.5 Programs for Machine Learning*. USA: Morgan Kaufmann Publishers, Inc.

Quintela, H. (2005). *Sistemas de Conhecimento Baseados em Data Mining: Aplicação à análise da Estabilidade de Estruturas Metálicas.* Dissertação de Mestrado, Universidade do Minho.

Ribeiro, A., Dinis, R., & Santos, M. (2011). *A Business Intelligence Infrastructure Supporting Respiratory Health Analysis*. Paper presented at the The First International Conference on Business Intelligence and Technology, Rome, Italy.

Santos, M. F., & Azevedo, C. (2005). *Data Mining e Descoberta de Conhecimento em Base de Dados*. Lisbon, Portugal: FCA.

Santos, M. Y., & Ramos, I. (2006). Como tornar o seu negócio realmente competitivo – Desafios tecnológicos e de gestão. *CXO: Tecnologias de Informação para Executivos, 1*(4), 56-61.

Santos, M. Y., & Ramos, I. (2009). *Business Intelligence - Tecnologias da Informação na Gestão de Conhecimento* (2 ed.). Lisbon, Portugal: FCA.

Sen, A. (2004). Metadata Management: Past, Present and Future. *Decision Support Systems, 37*(1).

SPP. (1997). *Normas clínicas para intervenção na Doença Pulmonar Obstrutiva Crónica*. Sociedade Portuguesa de Pneumologia Retrieved from http://www.sppneumologia.pt/.

Thuraisingham, B. (1999). Data Mining Technologies, Techniques Tools and Trends. *CRC Press LLC*.

Turban, E., Sharda, R., & Delen, D. (2010). *Decision support and business intelligence systems* (9 ed.). Upper Saddle River, NJ, USA: Prentice Hall Press

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). *Conceptual modeling for ETL processes.* Paper presented at the 5th ACM international workshop on Data Warehousing and OLAP, Virginia, USA.

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. Chichester, West Sussex, United Kingdom: John Wiley & Sons Inc.

Vivacare. (2010). COPD (Emphysema and Chronic Bronchitis), from http://fromyourdoctor.com/topic.do?t=7697

Wu, M. C., & Buchman, A. P. (1997). *Research Issues in Data Warehousing*. Paper presented at the BTW'97, Ulm, Germany.

Bibliografia

Burstein, F., & Holsapple, C. W. (2008). *Handbook on Decision Support Systems 1: basic themes*: Springer.

Burstein, F., & Holsapple, C. W. (2008). *Handbook on Decision Support Systems 2: variations*: Springer.

Chennakesava, R. A. (2008). *Fuzzy Logic And Neural Networks Basic Concepts and Application*: New Age International (P) Ltd. Publishers.

Colliat, G. (1996). OLAP, Relational, and Multidimensional Database Systems. *ACM Sigmod Record, 25*(3), 64-69.

Fayyad, U. (1997). Mining Databases: Towards Algorithms for Knowlede Discovery. *IEEE Computer Society Technical Committee on Data Engineering, 21*(1), 39-48.

Gangadharan, G. R., & Swami, S. N. (2005). Business intelligence systems: design and implementation strategies. In V. Luzar-Stiffler & V. H. Dobric (Eds.), *Proceedings of the 26th International Conference on Information Technology Interfaces* (pp. 139-144). Zagreb, Croatia: IEEE.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: what's next in business intelligence? In K. Davis & M. Ronthaler (Eds.), *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP* (pp. 1-6). Washington, DC, USA: ACM.

Guyodo, G., Blanc, I., Boulben, J. L., Lefevre, B., De Bels, F., & Garnier, R. (2010). *The French Toxic Exposure Surveillance System: Adaptation of a Business Intelligence System for Toxicovigilance*. Clinical Toxicology, 48(3).

Honghua, D., Ramakrishnan, S., & Chengqi, Z. (2004). *Advances in Knowledge Discovery and Data Mining: Proceedings of the 8th Pacific-Asia Conference, PAKDD 2004*. Sydney, Australia: Springer.

Horvath, M., Cozart, H., Ahmad, A., Langman, M. K., & Ferranti, J. (2009). *Sharing Adverse Drug Event Data Using Business Intelligence Technology*.Journal of Patient Safety,5(1),35-41.

Larson, B. (2006). *Delivering Business Intelligence with Microsoft SQL Server 2005*. Berkeley, CA, USA: Osborne/McGraw-Hill.

Larson, B. (2009). *Delivering Business Intelligence with Microsoft SQL Server 2008*: McGraw-Hill.

Lee, J. H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Elsevier Expert systems with applications, 29*(1), 145-152.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development, 2*(4), 314-319.

Mohammed, J. Z., Jeffrey, X., Ravindran, B., & Vikram, P. (2010). *Advances in Knowledge Discovery and Data Mining: Proceedings of the 14th Pacific-Asia Conference, PAKDD 2010 - Part 2*. Hyderabad, India: Springer.

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems, 13*, 177-195.

Negash, S., & Gray, P. (2008). Business Intelligence. In F. Burstein & C. W. Holsapple (Eds.), *Handbook on Decision Support Systems 2: variations*: Springer.

Pinto, F. (2005). *A descoberta de conhecimento em bases de dados como suporte a actividades de business intelligence : aplicação na área do database marketing.* Dissertação de Mestrado, Universidade do Minho.

Santos, M. Y., & Gonçalves, D. (2010). *Analysis of the Quality of Life after an Endoscopic Thoracic Sympathectomy: A Business Intelligence Approach*. Paper presented at the Second International Conference on Advances in Databases, Knowledge, and Data Applications.

Sarawagi, S., Agrawal, R., & Megiddo, N. (1998). Discovery-driven exploration of OLAP data cubes. *Advances in Database Technology—EDBT'98*, 168-182.

Thierauf, R. J. (2001). *Effective Business Intelligence Systems*. Westport, CT, USA: Greenwood Publishing Group.

Thomsen, E. (2002). *OLAP Solutions: Building Multidimensional Information Systems* (2 ed.): John Wiley & Sons, Inc.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *IEEE Computer, 40*(9), 96-99.

Anexos

Anexo A: *Script* para inserção automática de dados na tabela de factos

- InsertIntoFactTable.asp -

```
<%@ LANGUAGE="JAVASCRIPT"%>
<html>
<head>
<!--#include
file="ligacaoFactTable.inc"-->
<!--#include file="goTo.inc"-->
<link rel="stylesheet" type="text/css"
href="style.css" media="screen" />
</head>
<body>
<%
var CSFT = minhaLigacaoFactTable();
var CSDI = minhaLigacaoDadosIniciais();
var myConnectionDI =
Server.CreateObject("ADODB.Connection");
myConnectionDI.open(CSDI);
RSDI = myConnectionDI.Execute("Select *
from DadosIniciais");
var myConnectionFT =
Server.CreateObject("ADODB.Connection");
myConnectionFT.open(CSFT);
while(!RSDI.eof)
                        {
                        for(i=0; i <
(RSDI.fields.count); i++)
                                {
                                if (i==0)
                                        {
var idPatient = RSDI(i);
                                        }
                                if (i==3)
                                        {
var locality = RSDI(i);
RS3 = myConnectionFT.Execute("Select
IDLocation  from DimLocation where
Locality='"+locality+"'");
var idLocation = RS3(0);
                                        }
                                if (i==7)
                                        {

var smoker = RSDI(i);
var exSmoker = RSDI(i+1);
RS7 = myConnectionFT.Execute("Select
IDSmokeCharacterization  from
DimSmokeCharacterization where
Smoker='"+smoker+"' and
ExSmoker='"+exSmoker+"'");
var idSmokeCharacterization = RS7(0);
                                        }
                                if (i==9)
                                        {
var noseSneeze = RSDI(i);
var lacrimationItch = RSDI(i+1);
var allergies = RSDI(i+10);
var rhinitis = RSDI(i+11);
RS9 = myConnectionFT.Execute("Select
IDAllergyCharacterization  from
DimAllergyCharacterization where
NoseSneeze='"+noseSneeze+"' and
LacrimationItch='"+lacrimationItch+"'
and Allergies='"+allergies+"' and
Rhinitis='"+rhinitis+"'");
var idAllergyCharacterization = RS9(0);
                                        }
                                if (i==11)
                                        {
var dryCough = RSDI(11);
var coughALot = RSDI(12);
var dailyExpectoration = RSDI(13);
var wheezing = RSDI(14);
var wheezingTwelveMonths = RSDI(15);
var wheezingFlu = RSDI(16);
RS11 = myConnectionFT.Execute("Select
IDCoughCharacterization  from
DimCoughCharacterization where
DryCough='"+dryCough+"' and
CoughALot='"+coughALot+"' and
DailyExpectoration='"+dailyExpectoration
+"' and Wheezing='"+wheezing+"' and
WheezingTwelveMonths='"+wheezingTwelveMo
nths+"' and
WheezingFlu='"+wheezingFlu+"'");
var idCoughCharacterization = RS11(0);
                                        }
                                if (i==17)
                                        {
var moreFatigueThanPeopleSameAge =
RSDI(17);
var shortnessBreath = RSDI(18);
RS17 = myConnectionFT.Execute("Select
IDFatigueCharacterization  from
DimFatigueCharacterization where
MoreFatigueThanPeopleSameAge='"+moreFati
gueThanPeopleSameAge+"' and
ShortnessBreath='"+shortnessBreath+"'");
var idFatigueCharacterization = RS17(0);
                                        }
                                if (i==33)
                                        {
var FEV = RSDI(33);
var FEF2575 = RSDI(34);
        if(parseInt(FEV)>80)
```

```
        {
        var gravityLevel = 0;
        }
        if(parseInt(FEV)<=80 &&
parseInt(FEV)>70)
        {
        var gravityLevel = 1;
        }
        if(parseInt(FEV)<=70 &&
parseInt(FEV)>49)
        {
        var gravityLevel = 2;
        }
        if(parseInt(FEV)<=49 &&
parseInt(FEV)>29)
        {
        var gravityLevel = 3;
        }
        if(parseInt(FEV)<=29 &&
parseInt(FEV)>0)
        {
        var gravityLevel = 4;
        }
        if(parseInt(FEV)< 0)
        {
        var gravityLevel = -1;
        }
                                        }
                                if (i==21)
                                        {
var bronchialAsthma = RSDI(21);
var asthmaMedicationDaily = RSDI(22);
var asthmaMedicationCrisis = RSDI(23);
var cOPD = RSDI(24);
var fluMoreThanTwiceYear = RSDI(25);
var pneumonia = RSDI(26);
var pulmonaryTuberculosis = RSDI(27);
var tuberculosisAge = RSDI(28);
var bCGVaccine = RSDI(29);
var fluVaccineUsually = RSDI(30);
var pneumoniaVaccine = RSDI(31);
var
otherVaccinesPreventionRespiratoryInfect
ions = RSDI(32);
RS21 = myConnectionFT.Execute("Select
IDPulmonaryDiseasesCharacterization
from
DimPulmonaryDiseasesCharacterization
where
BronchialAsthma='"+bronchialAsthma+"'
and
AsthmaMedicationDaily='"+asthmaMedicatio
nDaily+"' and
AsthmaMedicationCrisis='"+asthmaMedicati
onCrisis+"' and COPD='"+cOPD+"' and
FluMoreThanTwiceYear='"+fluMoreThanTwice
Year+"' and Pneumonia='"+pneumonia+"'
and
PulmonaryTuberculosis='"+pulmonaryTuberc
ulosis+"' and
TuberculosisAge="+tuberculosisAge+" and
BCGVaccine='"+bCGVaccine+"' and
FluVaccineUsually='"+fluVaccineUsually+"
' and
PneumoniaVaccine='"+pneumoniaVaccine+"'
and
OtherVaccinesPreventionRespiratoryInfect
ions='"+otherVaccinesPreventionRespirato
ryInfections+"'");
var idPulmonaryDiseasesCharacterization
= RS21(0);
                                        }
                        }
                        RSF =
myConnectionFT.Execute("Insert INTO
FactFPP(Patient,IDTime,FEV1,FEF2575,Grav
ityLevel,IDSmokeCharacterization,IDPatie
nt,IDLocation,IDAllergyCharacterization,
IDCoughCharacterization,IDFatigueCharact
erization,IDPulmonaryDiseasesCharacteriz
ation) values (1, 1, "+FEV+",
"+FEF2575+", "+gravityLevel+",
"+idSmokeCharacterization+",
"+idPatient+", "+idLocation+",
"+idAllergyCharacterization+",
"+idCoughCharacterization+",
"+idFatigueCharacterization+",
"+idPulmonaryDiseasesCharacterization+")
");
                        RSDI.MoveNext()
                        }
myConnectionDI.close();
myConnectionFT.close();
%>
<script language="javascript">
alert('Dados GRAVADOS na Base de
Dados.');
</script>
</body>
</html>
```

- ligacaoFactTable.inc -

```
<%
function minhaLigacaoFactTable()
{
       return
"Provider=SQLOLEDB.1;uid=sa;password=***
***;Initial Catalog=FPPDM;Data
Source=INSYS2-309E4B0B\\MSSQLSERVER08";
}
function minhaLigacaoDadosIniciais()
{
       return
"Provider=SQLOLEDB.1;uid=sa;password=***
***;Initial Catalog=DadosFPP;Data
Source=INSYS2-309E4B0B\\MSSQLSERVER08";
}
%>
```

- goTo.inc -

```
<%
       function GoTo(url) {
              Response.Write('<script
language="javascript"
type="text/javascript">window.location =
"' + url + '";</script>');
       }
%>
```

Anexo B: *Query* de criação da tabela principal para análise com *Data Mining*

```
SELECT F.FEF2575,F.FEV1,F.SeverityStage, A.*, C.*, FA.*, L.*, P.*, PD.*, S.*
INTO TestTable
FROM FactFPP F
LEFT JOIN DimAllergyCharacterization A ON
F.IDAllergyCharacterization=A.IDAllergyCharacterization
LEFT JOIN DimCoughCharacterization C ON
F.IDCoughCharacterization=C.IDCoughCharacterization
LEFT JOIN DimFatigueCharacterization FA ON
F.IDFatigueCharacterization=FA.IDFatigueCharacterization
LEFT JOIN DimLocation L ON F.IDLocation=L.IDLocation
LEFT JOIN DimPatient P ON F.IDPatient=P.IDPatient
LEFT JOIN DimPulmonaryDiseasesCharacterization PD ON
F.IDPulmonaryDiseasesCharacterization=PD.IDPulmonaryDiseasesCharacterization
LEFT JOIN DimSmokeCharacterization S ON
F.IDSmokeCharacterization=S.IDSmokeCharacterization
where F.SeverityStage > 1
```